DOCUMENT RESUME

ED 150 191 TM 006 925

TITLE	Testing and the Public Interest.
INSTITUTION	Educational Testing Service, Princeton, N.J.
PUB DATE	77
NOTE	82p.; Proceedings of the Educational Testing Service Invitational Conference (37th, New York, New York, October 30, 1976)
AVAILABLE FROM	Invitational Conference, Educational Testing Service, Princeton, New Jersey 08541 ($5.00)
EDRS PRICE	MF-$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS	Achievement Tests; Awards; Competitive Selection; *Conference Reports; Criterion Referenced Tests; Cultural Disadvantagement; Culture Free Tests; Educational Assessment; Educational Testing; Elementary Secondary Education; Higher Education; Item Sampling; Low Ability Students; Minority Groups; *Norm Referenced Tests; Performance Factors; Racial Discrimination; Scores; Sex Discrimination; *Standardized Tests; *Student Attitudes; *Test Bias; Test Construction; Testing; *Testing Problems; Test Interpretation; Test Reliability; Test Validity
IDENTIFIERS	Tailored Testing; Test Anxiety; Test Theory

ABSTRACT

The 1976 Educational Testing Service (ETS) Invitational Conference served as a platform for individuals who have been prominent in educational measurement and research to present their views on issues surrounding the testing controversy. The 1976 ETS "The Testing Scene: Chaos and Controversy," presents a historical review of events surrounding the testing controversy. In "Test Theory and the Public Interest," Frederic M. Lord suggested three alternatives for solving some of the problems of cultural test bias: weighted scoring techniques, tailored testing, and item sampling. Esther E. Diamond discussed test construction techniques that could alleviate bias, in "Testing: The Baby and the Bath Water Are Still With Us." In "One Man's View of Testing," William Raspberry stated that norming contributes to cultural bias. Thelma T. Daley discussed the effects of testing on students, in "The Student and Testing." In the final paper, "Where Ignorance Is Bliss--'Tis Folly to be Testing," Robert L. Thorndike recommended that tests be evaluated in light of the decisions that will be based upon their results. (BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY
EDUCATIONAL TESTING SERVICE
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

ED150191
TM006 925

PROCEEDINGS OF THE 1976 ETS INVITATIONAL CONFERENCE

Testing and the public interest

Testing and the public interest

PROCEEDINGS OF THE 1976
ETS INVITATIONAL CONFERENCE

EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY

ATLANTA, GEORGIA
AUSTIN, TEXAS
BERKELEY, CALIFORNIA
EVANSTON, ILLINOIS
LOS ANGELES, CALIFORNIA
SAN JUAN, PUERTO RICO
WASHINGTON, D. C.
WELLESLEY HILLS, MASSACHUSETTS

The thirty-seventh ETS Invitational Conference, sponsored by Educational Testing Service, was held at the New York Hilton, New York City, on October 30, 1976.

Chairman: WILLIAM W. TURNBULL
President
Educational Testing Service

Educational Testing Service is an Equal Opportunity Employer.

Copyright © 1976, 1977 by Educational Testing Service. All rights reserved.

Library of Congress Catalog Number: 47-11220
Printed in the United States of America

Contents

v Foreword by William W. Turnbull

ix Presentation of 1976 Measurement Award to Ralph Winfred Tyler

Morning Session

3 The Testing Scene: Chaos and Controversy
Joan Bollenbacher

17 Test Theory and the Public Interest
Frederic M. Lord

31 Testing: The Baby *and* the Bath Water are Still With Us
Esther E. Diamond

Luncheon Address

47 One Man's View of Testing
William Raspberry

Afternoon Session

55 The Student and Testing
Thelma T. Daley

65 Where Ignorance is Bliss—
'Tis Folly to be Testing
Robert L. Thorndike

Foreword

Over the last few years, there has been widespread debate on various concerns and issues surrounding testing. The participants in the current debate include not only people with expertise in measurement or responsibility for giving tests and interpreting their results, but also the media, unions, ethnic groups, those who take the tests, professional associations, the courts, and the general public. Given the cross currents and contradictions, it seemed appropriate to provide a platform for individuals who have been prominent in the professional associations relating to educational measurement and research to present their views of the issues, the evidence with regard to them, and some possible ways to solve them.

The 1976 ETS Invitational Conference served as such a platform, and the speakers discussed issues relating to testing as well as some changes in testing practices. Their respective papers addressed past and present events in the testing scene; test theory in evaluation and design of tests; purposes of tests and ways in which test results are presented, interpreted, and used; aspects of testing and related practices that affect the student; and different types of decisions for which information provided by testing may be relevant.

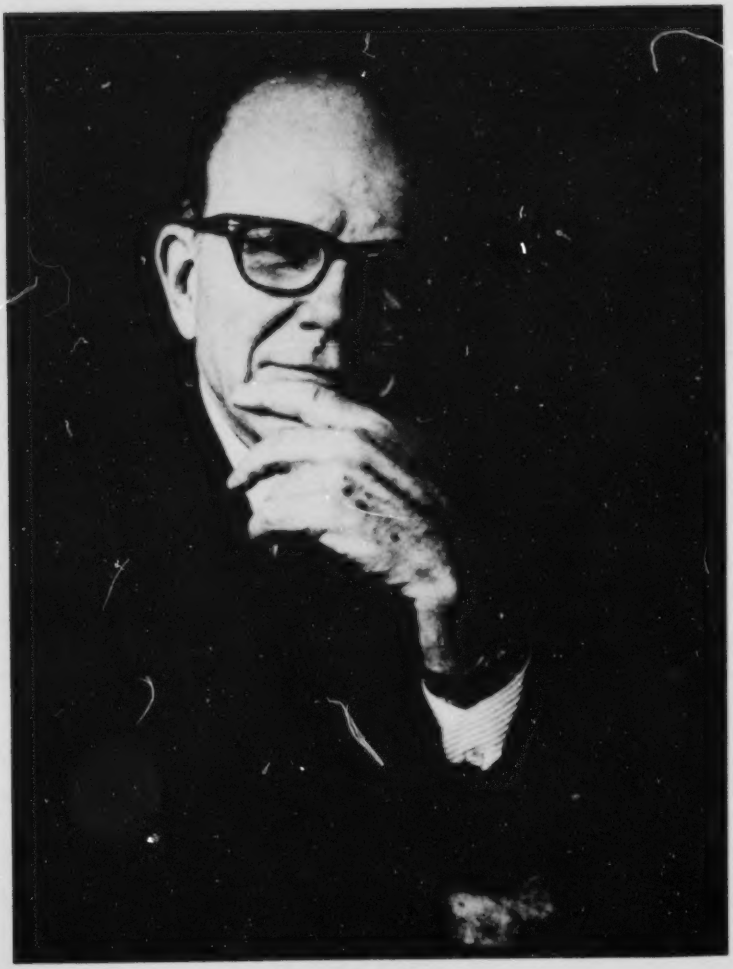
We are indebted to all of the speakers for sharing both their positive and critical views of the role of measurement in education and society. I should like to thank William Raspberry, a columnist at *The Washington Post*, for his candid luncheon speech in which he presented his views on the current attacks on standardized tests.

William W. Turnbull
PRESIDENT

EDUCATIONAL TESTING SERVICE

Measurement Award
1976

RALPH WINFRED TYLER



Ralph Winfred Tyler

The ETS Award for Distinguished Service to Measurement was established in 1970, to be presented annually to an individual whose work and career have had a major impact on developments in educational and psychological measurement. The 1976 Award was presented at the conference by ETS President William W. Turnbull to Dr. Ralph Winfred Tyler with this citation:

> For fully half a century Ralph Tyler has prodded education and educational measurement to become both more flexible and more focused, challenging us to conceptualize and assess those qualities that are hard to reach and hard to measure but are easily proclaimed as important goals of education. As Director of Evaluation of the monumental Eight-Year Study, he helped to shift education in this country from a narrow conception of subject-matter learning to a broader conception of growth and development of individuals, from a restrictive reliance on information, knowledge, and skills to an encompassing awareness of attitudes, appreciations, interests, and personal-social adaptability. By continuously emphasizing the functions of measurement in improving instruction, he helped to open both curriculum design and educational evaluation to a wide range of specific objectives and outcomes formerly lost in vague rhetoric.
>
> As creator and chief architect of the National Assessment of Educational Progress, he developed the financial, organizational, and political arrangements needed to make that massive and controversial concept into a practical and esteemed reality, while at the same time shaping its technical components to pioneer in the application of objectives-referenced measurement and criterion-referenced interpretation at the item level.
>
> As Director of the Center for Advanced Study in the Behavioral Sciences at Stanford, California, he fostered an atmosphere both challenging and supportive in which creative scholarship and interdisciplinary interplay flourished. There, during fifteen years as administrator, colleague, raconteur and wit, he personally influenced the development of hundreds of distinguished behavioral scientists.
>
> For his many contributions to the theory and practice of education, educational measurement and evaluation, and for his productive career as teacher and administrator, ETS is pleased to present the 1976 Award for Distinguished Service to Measurement to: Ralph Winfred Tyler.

Previous Recipients of the ETS Measurement Award

1970 E. F. Lindquist
1971 Lee J. Cronbach
1972 Robert L. Thorndike
1973 Oscar L. Buros
1974 J. P. Guilford
1975 Harold Gulliksen

Morning Session

The Testing Scene: Chaos and Controversy

JOAN BOLLENBACHER
Coordinator, Testing Services
Cincinnati Public Schools

When a scene is characterized by "chaos" and "controversy," it is reasonable to assume some past events have contributed to the disturbing situation. Since I have so described the current testing, I deem it appropriate to review some events of the past two decades that may help to put the current scene into perspective.

Before proceeding, however, I think it wise to explain how I conducted this review. As a result of administering a testing program in a large school system for a number of years, I have files of fat folders containing miscellaneous articles, some convention programs, and various publications that seemed important enough to survive several rounds of cleaning the files. In addition, I have several bookshelves of hardbacks and paperbacks pertaining to testing. This miscellaneous collection provided the sources for this review. Obviously, I make no claim for the completeness of the collection or of the review, but I hope you will agree that I have gleaned some interesting and, I hope, pertinent information.

At the outset, I believe it is appropriate to establish the date marking the beginning of the present testing controversy. I believe we can say it was not so very long ago, on October 4, 1957, the day the Russians sent into the sky a satellite called Sputnik. At first, the American people reacted with shock and disbelief that another nation appeared to be winning the space race. As soon as they tried to assess why our nation lagged behind, they immediately began to look critically at the quality and achievement of the schools. Within a year, Congress passed the National Defense Education Act (NDEA) which provided funds for many school systems to establish extensive testing programs. Accordingly, the administration of standardized tests expanded at a rapid rate.

That same year, 1958, a move by the National Merit Scholarship Corporation presented a problem to the schools. When the scholarship

program was inaugurated three years earlier. Testing was limited to the upper five percent of the seniors. The Scholarship Corporation changed its test publisher from Educational Testing Service (ETS) to Science Research Associates (SRA), moved from testing seniors in the fall to juniors in the spring, and suggested that the schools encourage many students to register for the test, even though they were not competing for scholarships. Soon thereafter, ETS began publishing the Preliminary Scholastic Aptitude Test (PSAT) which was offered to high school juniors in the fall. At their annual meeting in the spring of 1958, the secondary school principals protested the change because of the proliferation of external tests for high school students. The next year, Martin Essex, president of the American Association of School Administrators (AASA), appointed a nine-member committee to study the problems in testing and sent a questionnaire to school superintendents.

Meanwhile the technology had been developed for scoring and processing massive numbers of tests at what then was an almost unbelievable speed. Simultaneously, a second test for college admission, the American College Tests (ACT), had been developed and appeared in 1959, in time for use in what had come to be known as the "college admissions crisis of the sixties," which in turn was resulting from the post-World War II baby boom.

At their annual meeting in February 1960, the National Council on Measurement in Education (NCME)[20] and the American Educational Research Association (AERA) convened a symposium of five leading educators who addressed themselves to the topic "Resistance to Testing." The uses of aptitude and achievement tests were discussed, as well as the problem of who would be eliminated by such tests.

In March 1960, a test was given to 440,000 students in 1,353 secondary schools in all parts of the country. It was the comprehensive two-day battery of tests which was part of a large-scale, long-range research study known as Project TALENT. The study was being conducted by the American Institutes for Research and supported by funds from the U.S. Office of Education.

The U.S. Office of Education also became concerned about the increasing criticisms of tests as evidenced by their publication *Understanding Testing*[18], edited by Kenneth McLaughlin. The foreword by Lawrence Derthick, then Commissioner of Education, was in the form of an open letter to parents and teachers who were assured that Title V of NDEA was "in no sense a Federal testing program."

In quick succession, there appeared several paperbacks and hard-

backs relating to testing. There was the paperback entitled *Score, The Strategy of Taking Tests*[16] by Darrell Huff. A short time later came *Meeting the Test*[3], by Scarvia Anderson, Martin Katz, and Benjamin Shimberg. Then there was Banesh Hoffman's *The Tyranny of Testing*[13], Chauncey and Dobbin's *Testing: Its Place in Education Today*[4], and Gene Hawes' *Educational Testing for the Millions*[12].

While all the testing and discussion were going on in the high schools and colleges, the elementary school principals also had some questions with the result that two issues of the *National Elementary Principal* (September and November, 1961) were devoted to educational measurement—one to purposes and techniques and the other to interpretation and use. In contrast to the two recent issues of the *Principal* devoted to standardized testing, the 1961 issues featured a group of authors who would have comprised a "who's who" in the testing field.

In the meantime, there were increasing rumblings and grumblings by high school students and their parents about the numbers of tests required of candidates for college admissions and scholarship awards. Their protests resulted in the publication in 1962 of *Testing, Testing, Testing*[17], a 32-page paper-bound book prepared by a Joint Committee on Testing appointed by three national associations—the school administrators, the chief state school officers, and the secondary school principals. This book caused shock waves up and down the testing world. A few quotes will illustrate:

> The standardized test is, at best an ad hoc device; therefore, its function is limited. In comparison with the scope and duration of experiences to which a human being is subjected during his lifetime, the standardized test is a low-order hurdle.
>
> ... Like modern wonder drugs, standardized tests have captured the public mind.
>
> ... Most test makers are more or less candid about the limitations of standardized tests. But it is a mistake to assume that their knowledge and restraint have been appreciated by the public, or for that matter, even by many educators.

As I reread this little book I thought a lot of time and effort could have been saved if the critics of recent years had reprinted *Testing, Testing, Testing*. It condensed in 32 pages most of the criticisms contained in several lengthy recent publications.

I trust the foregoing list of events and publications provides enough evidence that criticism of tests is not a recent phenomenon. Now let us

consider several events of national importance which have had a significant effect on testing and added new dimensions to the criticisms beyond the problem of mere numbers of tests.

In 1964, Congress passed the Civil Rights Act. Subsequently a number of suits on the issue of segregation have been filed in the Federal Courts where testimony on standardized tests was involved. A federal agency, the Equal Employment Opportunity Commission (EEOC), was formed and issued a set of guidelines which set forth what an employer must do when using tests for selecting employees. Suits also have been filed in the Federal Courts on the issue of discrimination in the use of tests for employee selection. Possible race and sex bias in tests became serious concerns for the defense. The sex bias issue has been further emphasized by Title IX of the Educational Amendments of 1972.

In 1965, the year following the Civil Rights Act, Congress passed the Elementary and Secondary Education Act (ESEA). Title I of this act provided massive federal funding for the education of the disadvantaged. The evaluation requirements of Title I involved extensive use of standardized reading tests, with the result that some school children were subjected to massive overtesting. Nine years later, in 1974, the *Anchor Test Study*[27], which involved equating eight standardized reading tests, was a direct result of the Title I evaluation problems.

Back in 1966, the *Equality of Educational Opportunity Study*[26], known as the Coleman Report, was published. Many of its conclusions, which profoundly affected the public schools, were based on the results of standardized achievement tests.

Meanwhile, as these national events were taking place, back at the American Psychological Association (APA) a committee composed of eight members from APA, AERA, and NCME completed three years of work and published the *Standards for Educational and Psychological Tests and Manuals*[1] in 1966.

By 1967, planning for the National Assessment of Educational Progress (NAEP) had been under way for three years, but early that year school administrators registered serious objections. Most magazines and newspapers had articles on the subject, with the *New York Times* of February 12 calling National Assessment "one of the most hotly contested issues in American education."

That same year, the College Entrance Examination Board (CEEB) appointed a 21-member Commission on Tests charged to review the College Board's testing functions, to consider possibilities for funda-

mental changes in tests and their use, and to make recommendations accordingly. The Commission's report[5] was issued three years later, in 1970.

Back now to May 25, 1969, the date of the Conference on the Ethical and Legal Aspects of School Record Keeping convened by the Russell Sage Foundation. The report of the Conference resulted in the publication of a set of "Guidelines for the Collection, Maintenance and Dissemination of Pupil Records," which in turn provided basic information for the Family Education Rights and Privacy Act, known as the Buckley Amendment, passed by Congress in 1974. No longer could test scores be kept secret from students and parents.

Now let's take another look at the late 60s, a time of student rebellion in the colleges and universities which was reflected in the schools. Teachers reported that students sometimes resented tests, sometimes resisted them, and sometimes rebelled against them. Just as schools were trying to cope with these changing attitudes, accountability reared its ugly head. By 1970, it was hard to find any educational conference that did not have at least one session on accountability. At most of them, there was a discussion of the uses and limitations of standardized tests in meeting accountability demands. Those who predicted trouble ahead were correct.

On Valentine's Day in 1971, the *New York Times* reported "in a historic move the (New York) board (of education) announced that it would establish procedures to hold the schools and their staffs accountable for their success in educating children." The *New York Times* does not use lightly terms like "historic move," even on Valentine's Day. The article reported that the move was supported by Albert Shanker of the American Federation of Teachers (AFT). Those who follow events in the New York schools will be interested in the report, "Security in a Citywide Testing Program," by Anthony J. Polemeni, published by the National Council on Measurement in Education[24].

Just a year after the *New York Times* article, 650 members of the National Education Association (NEA) who met at the annual NEA Conference on Civil and Human Rights called for an immediate moratorium on standardized testing. There are those who would say that from there on it has been downhill all the way.

At the time the NEA was calling for a moratorium, APA, AERA, and NCME were working on the revision of the 1966 *Standards*. A section on "Standards for the Use of Tests" was added to the publication. After

several drafts of the document had been completed, the NEA was asked for its reactions. The NEA representatives felt they were being asked after the fact, and they declined to participate. The revised *Standards*[2] were published in 1974. I fear, however, that the document in its present form has not had wide circulation beyond psychologists and students in classes in educational measurement. As I understand it, some translations are under way which may help.

Now I would like to comment on a few events of the past eighteen months which, in my judgment, have made it almost impossible for persons in the schools who have responsibilities for testing to cope with the resulting chaos and confusion. For openers, there was the March/April 1975 issue of the *National Elementary Principal*, the official publication of the National Elementary Principals Association. The cover was captioned "IQ, The Myth of Measurability." Most of the 16 articles were negative. Paul Houts, editor of the magazine, called for "an intensive national inquiry into standardized testing"[14]. The July/August issue[15] was devoted to a devastating attack on standardized testing, as well as a blast at the National Assessment of Educational Progress as an assessment that "asks powerless communities to assess themselves in terms provided by the powerful." The lead editorial stated that "...it is now imperative for the education profession to take the initiative in developing alternatives to the current tests. Testing must be returned to the education profession itself." The editor also called for immediate cessation of the practice of releasing test scores to the press.

The September/October issue of the magazine contained four letters to the editor approving the "IQ issue," but one letter from Professor Herbert Rudman[25] of Michigan State University registered violent exception. Regarding the contributors to the issue, he said, "We had professors of physics, animal conditioning, mathematics and the like. Nowhere did I find an author whose special competency, training, and experience qualified him to address as complex an issue as standardized testing."

Between the publication of the two issues of the *National Elementary Principal*, there appeared a new critic of the tests, the consumer advocate. In the May 1975 *Ladies Home Journal*, of all publications, there was an article in which Ralph Nader called for citizens whose lives are "shaped by the power of ETS to call to account the testers and the institutions that support them."

Before most elementary school principals had had time to read their

magazine, an interesting event occurred in Akron, Ohio, in late September. The Akron Board of Education was denied a motion for a new trial, and the Akron *Beacon Journal* won the suit forcing the Akron Board to release school-by-school test results to the press; therefore, the position of the editor of the *National Elementary Principal* relative to the release of test scores was not upheld by the court.

Another event in September 1975 did not help matters in the schools. The College Board reported that the average scores attained by the 1975 high school graduates on the Scholastic Aptitude Test (SAT) were the lowest ever. It was noted also that more women than men had taken the test. These declining scores on the SAT and also on the American College Test (ACT) had been a matter of continuing concern, to the extent that earlier the National Institute of Education (NIE) had called 30 experts from test organizations and colleges to Washington to try to define the problem, but with little success. The College Board also appointed a so-called "blue ribbon panel" to study the problem.

I cannot resist mentioning an article which appeared this past February in the Cincinnati Enquirer. It quoted Leo Munday, vice president of the American College Testing Program, as saying that the decline of college admission test scores may be partly due to an increase in "mediocre, college-bound female students." Mediocre, indeed!

Now we come to October 4, 1975. James J. Kilpatrick in his syndicated column commented on the issue of the *National Elementary Principal* devoted to standardized tests. He concluded:

> For a variety of reasons, public education is in deep trouble in America. We need urgently to know the dimensions of this trouble; we need to know which approaches, techniques and devices work and which ones fail. The innocent pupils can't tell us; the defensive educators don't want their schools compared; parents are ill-equipped for evaluation. That leaves the standardized tests. Defective as they are, we had better keep them in use.

Exactly one week later, October 11, 1975, Mr. William Raspberry of *The Washington Post* devoted his column to a discussion of the same issue of the *National Elementary Principal*. He concluded:

> Teachers (and school districts) who want to conceal how effectual they are, can avoid comparisons with other school units serving similar populations by avoiding standardized testing.
>
> I suspect that one of the reasons parents are reluctant to let go of standardized tests, as bad as they are, is that they don't trust the schools to give them candid evaluations of how well the schools are performing.

Meanwhile *Education USA*[7] reported that the annual evaluation of NAEP was conducted by a nine-member team appointed by the Department of Health, Education and Welfare. The team said that National Assessment's nonanalytical data are of limited use to states and schools. Then it was reported that the director of NAEP commented that he was "extremely optimistic that (they would) be able to provide information that will be used in the decision-making process in schools across the country." This year it was reported that a spokesperson (their term) for National Assessment responded to a General Accounting Office criticism saying that it is not NAEP's business to set national standards for test performance.

Such an exchange of comments can only be mindboggling to the teacher who might look upon National Assessment as criterion-referenced, only to learn that now it is suggested that it be norm-referenced. Add to this the article by James Popham[24] in the May 1976 *Phi Delta Kappan* suggesting that there can be normative data for criterion-referenced tests!

Now we come to November 1975, when representatives of some 35 or 40 national educational associations, government agencies and education groups met in Washington to consider implications of widespread use of standardized tests. The conference was convened by the National Association of Elementary School Principals and the North Dakota Study Group on Evaluation under a grant from the Rockefeller Brothers Fund. The next month the draft of a nine-item position statement[19] was released. Following the second meeting of the group in May, an *Education USA*[9] headline stated "Standardized Testing Issue Becoming Free-for-All" and reported that the symposium had not yet agreed on a basic statement about tests but that the participants "did square off at representatives of seven test publishing companies who attended." The third meeting of the group was held in the early fall of 1976, but as yet no agreement has been reached.

As if all of this controversy is not enough, even the National Council of Teachers of English (NCTE) added to the confusion. At their annual meeting last Thanksgiving in San Diego, the teachers defeated a resolution to eliminate sexist language from tests because they were afraid if they did pass a resolution, it would appear they favored standardized tests!

About the time English teachers were not considering test bias, the National Institute of Education (NIE) convened a three-day conference on bias in achievement tests. One account[4] reported that Robert Ebel

of Michigan State University said there is no direct evidence that achievement tests commonly used in this country are biased, and the chances of turning up such evidence are "quite improbable." Robert Green, also from Michigan State, disagreed, saying that the tests are inappropriate for many black, Spanish-speaking, and poor white families, and worst of all, low scores on such tests are used as an excuse for a watered-down curriculum. Dr. Green reportedly said, however, that he favored "cleaning up," not abolishing standardized tests, because, he said, "I'd sooner challenge the bias of tests than the biases of teachers and principals."

Another testing issue of major significance relates to the opposing viewpoints of the two major teacher organizations. The NEA position was widely publicized last February when Terry Herndon, NEA executive director spoke to the Commonwealth Club of San Francisco. The headline in the *NEA Reporter*[9] proclaimed, "Standardized Tests Must Go, Herndon Says." Conversely, the American Federation of Teachers (AFT) passed a resolution at its annual meeting in August 1976, indicating that instead of eliminating standardized tests, they should be improved, but they should not be used for evaluating teachers or staff performance.

While the arguments over standardized tests go on, a trend in the country which undoubtedly will involve considerably more testing should not be ignored—that is the back-to-the-basics movement. It was reported recently[10] that already five states have enacted minimal competency testing and 13 states have initiated studies or mandates on competencies. Criterion-based or not, that will be a lot of testing.

In a recent Gallup poll[11] the question was asked, "Should all high school students in the United States be required to pass a standard examination in order to get a high school diploma?" A total of 65 percent of the respondents answered "yes."

In reporting on the results of the same Gallup poll, a large headline in the September 25 issue of the Cincinnati *Enquirer* stated, "Americans Trust Standard Testing." When asked for reasons to explain the decline in national test scores, only 16 percent of those polled by Gallup gave as a reason that the tests are not reliable. Since we do have an interested public, it seems especially appropriate for us to consider the public interest!

Now, where does all of this lead us? Obviously, when the Elementary Principals Association, the National Council of Teachers of English and the NEA are objecting to standardized tests, there is a problem. In this

whole confusing business it seems to me that as educators we have an obligation to make some thoughtful recommendations regarding standardized testing rather than to make sweeping statements that "standardized tests must go." I cannot imagine suggesting that the test of General Educational Development (GED) be abolished, a test which is given annually to half a million adults across the United States to establish high school equivalency. We have given the test in Cincinnati for 30 years, and it alone has provided a passport to a better job and a better life for many persons in our city; yet *it* is a standardized test.

The college admissions tests (SAT and ACT) are other examples of standardized tests which should be considered. If we cast these out, are we going to return to the days when candidates for admission to college had to take a different placement test for each college where they applied? Previous to the SAT, the prestigious eastern colleges admitted students primarily from eastern prep schools and a few public schools. After the SAT was established, admissions officers discovered that there are capable pupils all over the country, and consequently student populations were drawn from a wider, more representative area. Also, before we throw out the SAT and ACT, we should think about the effects of the Family Rights and Privacy Act which opens all records to parents and students. Consequently, counselors and teachers are likely to be far less candid in their letters of recommendation. If we eliminate tests and letters of recommendation, the admissions officer has only grades and rank in class remaining. Class rank and school grades vary considerably from one school to another. Then, for the admissions officer who has *only* grades and class rank for decisions, it will be natural to favor the schools he or she knows best, and we are right back where we started.

If "standardized tests must go" means *no* testing, I think that is an unrealistic point of view in the society in which we live. People are selected continually on a variety of criteria for a variety of purposes. Teachers decide who will be promoted; admissions officers decide who will be admitted; employers decide who will get the job; football scouts decide who will get the scholarship; and baseball managers decide who will make the team. "Aha!" you may say "But many of their measures are criterion-referenced." True. But believe me, they are norm-referenced too. It is strange that no one objects to all of those norms in the world of athletics—the yards he has gained, how fast he runs, his batting average. Those are all compared against the performance of other

individuals. Maybe they are accepted because they are not called "standardized."

There is still another reason why it seems to be unrealistic to say that standardized tests can be eliminated. With the national average of over $1,000 a year as the cost to educate a pupil, it seems likely that the parents and taxpayers are going to want some evidence other than verbal assurance that the children are learning something.

And that leads me to still another point. I firmly believe that most teachers are working hard at teaching. If the children are not learning, then we need a good deal of evidence to establish why they are not learning. I am not talking about opinions, but what we call hard data. What about their attitudes? What about attendance and mobility? What about actual time devoted to instruction? And what about their achievement? It seems to me we have an obligation to help teachers with techniques for gathering and interpreting such data.

In the climate of the 70s we must consider standardized achievement tests as one of many kinds of data we use. They do provide good information about the achievement of individual pupils, especially when scores are collected on a longitudinal basis and especially when item data are provided for individuals and their classes. If we are to be fair, I think, too, we must recognize that reporting achievement data, at least for a school system, can provide important clues regarding improvement in instruction from year to year.

But I am afraid that those of us who work with tests and testing have not done our best to help teachers, principals, and the public to understand the strengths and limitations of standardized tests. Our neglect is illustrated in a statement in a booklet published by the National School Public Relations Association entitled *Releasing Test Scores: Educational Assessment Programs, How to Tell the Public.* Here is what it says:

> *Beware of Statisticians.* The natural impulse in attacking such a problem is to assemble the test specialists and statisticians to explain. *But beware of this seemingly simple approach.* Statisticians and test specialists enjoy talking to one another, and they do that very well. But they have trouble with educators. The evidence: Everything goes well until somebody asks a question. It's all downhill from there.
>
> If you have a test specialist or statistician on your staff who can popularize the presentation, you are in proximity to a rare jewel. If not, have them work very closely with your information specialists as they prepare their explanations.[22]

If this statement does not convince you that we have a problem, then I refer you to a statement by Henry Dyer when he spoke to the test directors of large school systems last May. Dr. Dyer summarized the issue succinctly, as he always does. He said:

> I find disturbing... the behavior of many psychologists, psychometricians, and other social scientists who find educational and psychological measurement a fascinating field of inquiry, but who retreat from all the controversies over testing and evaluation by retiring into cozy little coteries where they write beautiful essays to one another that are so heavily laced with mathematical equations that it is a rare person out there in the schools who can understand what they are talking about. Much of what they produce can be of extraordinary importance to your evaluator on the front line, but it is almost always buried so deep in technical books and journals that, for all intents and purposes, it is irretrievable.[6]

As an example, Dr. Dyer cited the *Journal of Educational Measurement* (JEM) published by the *National Council on Measurement in Education* (NCME). The irony is that NCME is intended to serve the practitioner. Lest some in the audience are concerned that I am suggesting JEM has no place in NCME, I wish to assure you that is farthest from my mind. What I am suggesting is that technical information be translated into publications that can be understood by those who are not psychometricians and measurement experts. A long time ago, AERA published a series called "What Research Says to the Teacher," but I know of no similar recent efforts.

By now you must have enough of chaos and controversy. Perhaps you may think that this recital of events in testing over two decades is a bit too much. I shall now conclude with one comment.

Publications about tests and testing are almost totally lacking in humor. As a Cincinnatian, I think it appropriate to say that 99-44/100 percent of them can be so categorized. Lest you think this paper provides ample proof that there is no humor in testing, I decided to take a drastic step to improve the situation and quote Art Buchwald, who was recently interviewed on the "Today" show. He was asked if the lack of humor in the presidential campaign presented a problem to him as a political satirist. He replied, "Just because there's no humor doesn't mean it isn't funny!"

References

1. American Psychological Association. *Standards for educational and psychological tests and manuals.* Washington, D.C.: American Psychological Association, 1966.
2. American Psychological Association. *Standards for educational and psychological tests.* Washington, D.C.: American Psychological Association, 1974.
3. Anderson, S., et al. *Meeting the test.* New York: Scholastic Magazines, Inc., 1963.
4. Chauncey, H., and Dobbin, J. *Testing: Its place in education today.* New York: Harper and Row, 1963.
5. College Entrance Examination Board. *Report of the Commission on Tests.* Two volumes. New York: College Entrance Examination Board, 1970.
6. Dyer, H. S. *Standardized tests: the current controversy in perspective.* Paper presented at the Large School Systems Invitational Conference on Measurement in Education, Indianapolis, May 1976.
7. Education USA, October 13, 1975, P. 39.
8. Education USA, January 5, 1976, P. 107.
9. Education USA, June 14, 1976, P. 247.
10. Education USA, September 6, 1976, P. 4.
11. Gallup, G. H. Eighth annual Gallup poll of the public's attitudes toward the public schools. *Phi Delta Kappan,* 1976, *58* (2), 187-200.
12. Hawes, G. R. *Educational testing for the millions.* New York: McGraw-Hill, 1964.
13. Hoffman, B. *The tyranny of testing.* New York: Crowell-Collier Press, 962.
14. Houts, P. L. Standardized testing in America. *National Elementary Principal,* 1975, *54* (4), 2-3.
15. Houts, P. L. Standardized testing in America, II. *National Elementary Principal,* 1975, *54* (6), 2-3.
16. Huff, D. *Score, the strategy of taking tests.* New York: Appleton-Century-Crofts, 1961.
17. Joint Committee on Testing. *Testing, testing, testing.* Washington, D.C.: American Association of School Administrators, 1962.
18. McLaughlin, K. F. *Understanding testing.* Washington, D.C.: U.S. Government Printing Office, 1960.

19. National Association of Elementary School Principals. *Testing group issues strong statement. Spectator*, December 1975. P. 4.

20. National Council on Measurement in Education. *Convention Program*, February 14, 1963. P. 6.

21. National Education Association. *NEA Reporter*, February 1976. P. 12.

22. National School Public Relations Association. *Releasing test scores*. Arlington, Va.: National School Public Relations Association, 1976. P. 43.

23. Polemeni, A. J. *Security in a citywide testing program*. National Council on Measurement in Education. *Measurement in education*. Summer 1975.

24. Popham, W. J. Normative data for criterion referenced tests? *Phi Delta Kappan*, 1976, 57 (9), 593-594.

25. Rudman, H. C. Letter to the editor. *National Elementary Principal*, 1975, 55 (1), 3.

26. U.S. Department of Health, Education, and Welfare. *Equality of educational opportunity*. Washington, D.C.: U.S. Government Printing Office, 1966.

27. U.S. Department of Health, Education, and Welfare. *Anchor test study*. U.S. Government Printing Office, 1974.

Test Theory and the Public Interest*

FREDERIC M. LORD
Senior Research Psychologist
Educational Testing Service

I am going to talk about several applications of test theory in the public interest. The thread running through the various applications is the evaluation and design of tests for particular individuals, or for particular subgroups, or at particular ability levels.

An outstanding recent application, not yet completely digested by psychometric experts, stems from three 1971 articles on test bias and culture fairness by Robert L. Thorndike[1], by Richard Darlington[2], and by Robert L. Linn and Charles Werts[3]. Until these articles appeared, many of us thought that we could determine whether a test or selection procedure was fair or unfair to minority groups by using a simple statistical procedure. One of Thorndike's important contributions was to make clear that what is fair according to one definition may be quite unfair according to another definition.

Consider the selection and hiring of job applicants, or the selection of people for admission to college. Suppose first of all that in advance of selection we have available some adequate criterion measure on all applicants. In this very unlikely situation, we might simply select the applicants on the basis of criterion score, regardless of group membership.

Whether or not this is a proper policy is a social question, not a measurement problem. The measurement problem arises whenever, as is ordinarily the case, we do not have the criterion measure available at the time of selection. Instead we have available a test score whose only virtue is that it predicts the criterion measure. The correlation

*Part of this talk and Figs. 3-6 are taken from a forthcoming paper in *Journal of Educational Measurement* titled "Practical Applications of Item Characteristic Curve Theory". Figs. 1-2 are taken from F. M. Lord, "Quick Estimates of Relative Efficiency of Two Tests as a Function of Ability Level," *Journal of Educational Measurement*, 1974, *11*, 247-254. Used by Permission.

between predictor and criterion is usually not very high, probably no higher than .60. What is the effect of selecting people on the basis of their criterion score?

A common selection procedure uses a single cutting score on the predictor without regard to group membership. This procedure maximizes the expected criterion score of the selected individuals. This seems eminently fair to the selecting institution, but is it fair to the individuals involved? And in particular to members of minority groups?

Suppose we could select on criterion score. Suppose that selecting on the criterion would result in selecting 50 percent, say, of all applicants from a certain minority group. Consider now the effect of substituting a predictor for the criterion score. It could happen that when we use a single cutting score on the predictor for selection, only 25 percent, say, of the minority group will be chosen. Such a result certainly does not seem fair to the minority group.

The selection procedure is still fair to the institution doing the selection. This institution will hire or admit the individuals with the highest expected criterion scores. But the use of predictor has clearly injured the minority group. Only half as many of this group will be selected as would be the case if the criterion were available at the time of selection. This is a major point made by Thorndike in his paper.

It seems clear that such a situation is a bad one. There are two possible approaches to correcting it. One approach, which has led to important papers by prominent workers in the field, is to try to correct the inequities resulting from a biased predictor by setting different cutting scores for different groups. The main conclusion from reading the papers on this subject seems to be that different sets of cutting scores will be utilized, and judged fair, by different people. There does not seem to be any way of correcting for a biased predictor in a way that will seem fair according to all reasonable value systems.

An alternative possibility, which is also being attempted, is to try to improve the predictor so that the same cutting score can be used for everyone. Whether or not a particular predictor is seriously unfair to some minority group depends on what the predictor measures. If the predictor measures some trait that is irrelevant for success, a minority group that happens to rank low on this irrelevant trait will obviously be unfairly treated by use of a single cutting score on this predictor. Again, if the predictor does not measure some trait that is important for success, a minority group that happens to rank high on this important trait will be unfairly treated by use of a single cutting score on this

deficient predictor. An obvious course is to try to improve our predictors so as to avoid the unfair situations.

It is interesting to ask what would happen if we could build a predictor that differed from the criterion only because of random errors of measurement. Would the use of such a predictor with a single cutting score still be unfair to minority groups? The answer by Linn and Werts is that such a procedure will slightly favor low-scoring groups and handicap high-scoring groups. The reason is that a predictor containing random errors of measurement will differentiate high-level and low-level groups less well than would the criterion score, were it available. This means that more people will be selected from low groups and fewer people from high groups.

This becomes particularly obvious in the extreme where the predictor is almost completely unreliable. If the predictor had zero reliability, it could not discriminate between one group and another group, which means that any two groups would have the same distribution of predictor scores. In such a case, clearly, use of a single cutting score on the predictor favors any group that is low on criterion score.

It may not be possible in many cases to produce mental tests that differ from an important criterion only because of errors of measurement. We certainly can work toward this, however. We can try to avoid predictors that measure some irrelevant trait, to the disadvantage of a minority group. If we cannot avoid using such predictors, then indeed we will have a difficult task deciding how to select cutting scores to compensate for measuring the wrong traits.

Let me now turn to a different subject. In classical test theory, the value of a test is usually summarized by one or more of three coefficients: the validity coefficient, the reliability coefficient, and the standard error of measurement. Any such coefficient describes the average performance on the test for a certain group.

The magnitude of the first two coefficients varies from group to group. In general, such a coefficient, reported by the publisher for a supposedly nationally representative group, will not be appropriate for any particular teacher and his or her class of students. A particular classroom is likely to have a smaller range of talent than a nationally representative group.

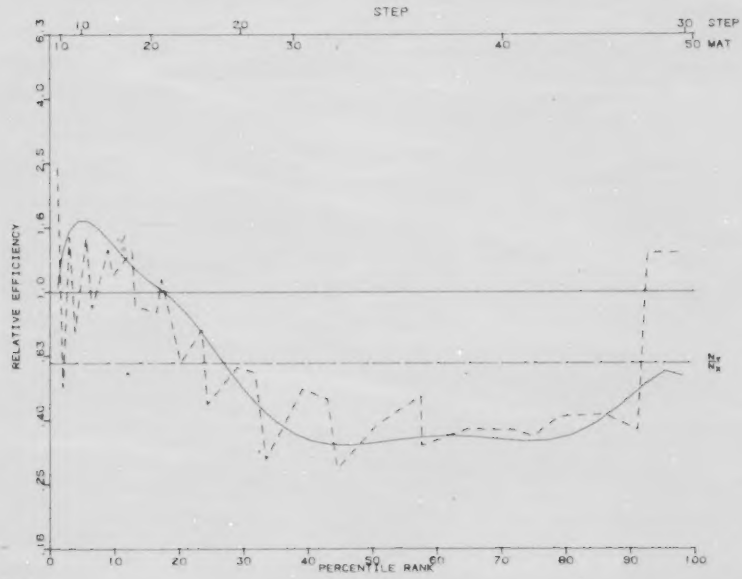
The standard error of measurement of a test may be reasonably constant from group to group, provided the groups are not very different in ability level. But now we have a different problem: we can compare standard errors of measurement from group to group, but not from test

to test. The standard error of measurement is expressed in terms of the raw score scale, which varies from one test to another. If we use standardized scores instead of raw scores, then we cannot compare standard errors of measurement from group to group.

What is needed is a method of describing the effectiveness of a test in a way that will be appropriate both for across-group comparisons and for across-test comparisons, provided that the tests are all measures of the same trait, ability, or skill. Does this sound impossible? We can come close to doing this.

Figure 1 shows the relative efficiency of two widely used tests of reading vocabulary. The relative efficiency varies according to level of developed ability, which is shown along the base line of the figure. Specifically, the figure shows the relative efficiency of a reading vocabulary score from the Sequential Tests of Educational Progress (STEP), relative to a reading vocabulary score from the Metropolitan Achievement Test (MAT). The data describe a particular form of each test.

Figure 1.
Relative efficiency of STEP compared to MAT.



20

What is meant by relative efficiency here? The efficiency of a single test at a particular ability level is inversely proportional to the squared standard error of measurement for people at that ability level.

If two tests measure on the same score scale, then their relative efficiency at a particular ability level is simply the ratio of their squared standard errors of measurement at that level. Since two tests from different publishers typically measure on different score scales, even though they are tests of the same ability, an adjustment must be made for differences in score scale. Thus the relative efficiency of one test with respect to another at a particular ability level is simply the ratio of their squared standard errors of measurement at that level adjusted for differences in score scale. If one test has a relative efficiency of .5 with respect to another at some ability level, then doubling the length of the first test will make it as efficient as the second test.

Figure 1 shows that the STEP test is more efficient than the MAT test at low ability levels, but less efficient at all other levels. This reflects the fact that the STEP test is much easier than the MAT test. It is well known that an easy test discriminates best among low-level students. A hard test discriminates best among high-level students.

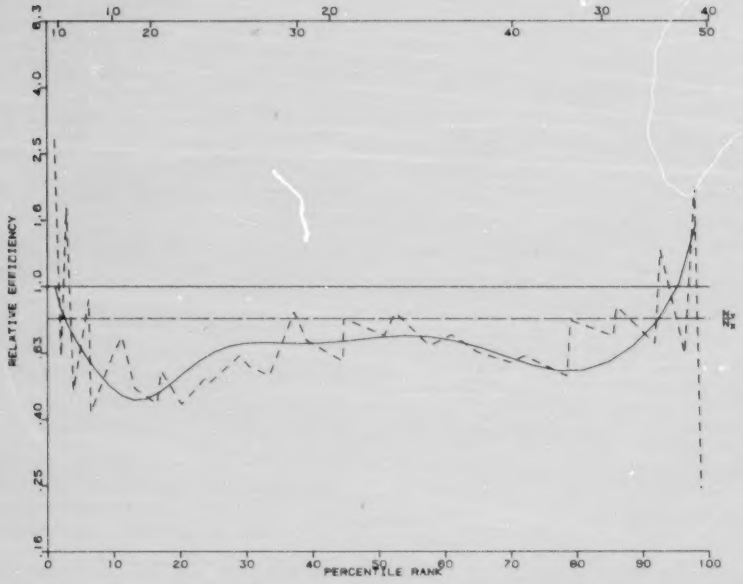
The STEP test is shorter than the MAT test. The dashed horizontal line shows the relative efficiency that would be found if the two tests differed only in length.

Figure 2 shows the relative efficiency of a particular form of another published reading vocabulary test compared to MAT. This test is less effective than MAT for most of the range of interest here.

In these figures the base line is calibrated in terms of percentile rank for a particular group of students. The top horizontal line is calibrated in terms of raw scores on both the tests administered. With the aid of such figures, if a teacher knows the ability level of his group or the ability levels at which he wishes to make effective discrimination, then he can make an informed choice among available published tests. This is much better than relying on coefficients reported by the publishers for groups that contain students at ability levels not relevant for this teacher.

How do we get these relative efficiency curves? They can be produced by a rather complicated and expensive process based on the estimation of item parameters by item response theory. Fortunately a usable approximation to the relative efficiency curves can be obtained directly from frequency distributions of number-right scores, as I have pointed out in a 1974 issue of the *Journal of Educational Measurement*[4]. The dashed jagged lines in the figures show the approximations obtained

Figure 2.
Relative efficiency of Form A, Reading Vocabulary Test compared to MAT.

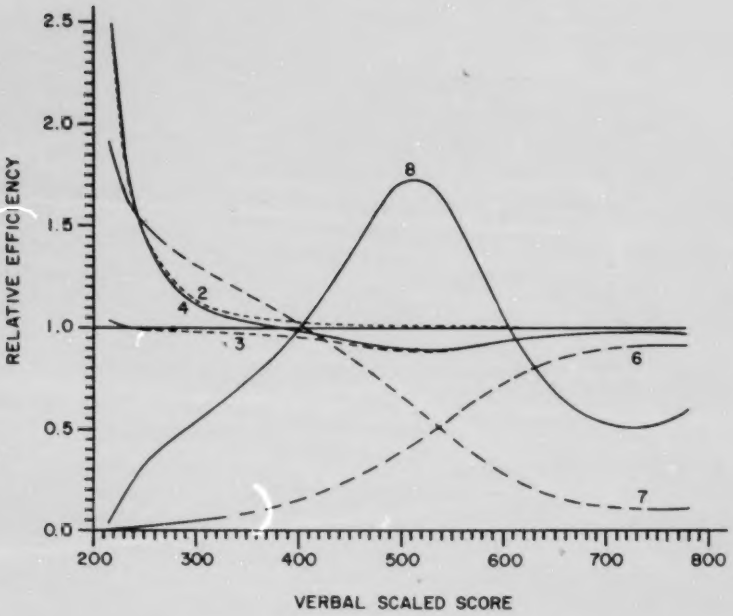


directly from the number-right score distributions, with the help of a desk calculator.

Such relative efficiency curves have many uses besides choosing among published tests. Recently at Educational Testing Service (ETS) and at the College Entrance Examination Board certain revisions of the *Scholastic Aptitude Test* (SAT) were contemplated. A possibly desirable revision was to try to make the tests easier for low ability students—provided this could be done without impairing the measurement effectiveness of the test for high ability students. It was decided to investigate the effects of various possible changes from existing forms of the test.

A particular form of the verbal SAT was chosen and analyzed. We then asked such questions as the following: Suppose we took the five easiest items in this form of the verbal SAT and added five more items with statistical properties exactly like these. What would be the relative efficiency of the resulting test? This relative efficiency, relative to the form of the test in actual use, is shown by curve 2 in Figure 3. As might

Figure 3.
Relative efficiency of various modified SAT Verbal tests.



be expected, the effectiveness of the test is slightly improved for examinees at low ability levels without much change in the effectiveness of the test elsewhere.

Curve 3 shows the effect of eliminating a block of five medium difficulty items in the middle of the test. Efficiency is impaired for middle ability students, but there is not too much effect elsewhere.

If we simultaneously add five easy items, as already described, and eliminate five items of medium difficulty, the relative efficiency of the resulting test is shown by curve 4. This is seen to be a sort of combination of the other two curves. It does seem to be possible to improve the measurement effectiveness of the test at low ability levels without sacrificing its effectiveness at high ability levels. However, we do lose effectiveness at medium ability levels. In general, experience shows that any gain achieved at one ability level is usually paid for by a loss of effectiveness at some other level. Usually the only way to avoid this

rule would be to write better items; but this increases the cost of test production.

There is something to be learned from curves 6, 7, and 8. Curve 6 shows what would happen if we simply discarded the easiest half of the items in the test. The half-length test would be almost as good as the full-length test for high-ability students. Such a test would of course be virtually useless for low-ability students. This tells us that the easiest half of the items in the current form of the SAT Verbal test are contributing very little towards measuring the high-ability students. In effect, only half the time spent by the high-ability students in taking the test is of any use for measuring them.

Curve 7 leads to a particularly interesting conclusion. Curve 7 represents the relative efficiency of a half-length test obtained by discarding the hardest half of the items in the Verbal SAT. In contrast to curve 6, notice that here throwing away half the items improves the measurement at low-ability levels. The reason is that low-ability examinees guess at random on hard items. The resulting random noise tends to drown out whatever measurement would otherwise be accomplished by the easier items.

The conclusion that I want to emphasize is that we cannot make a test appropriate for low-ability examinees simply by adding some easy items. As long as the test contains many hard items on which these examinees guess at random, the test cannot be a really effective measuring instrument for them.

Curve 8 shows the relative efficiency of a full-length Verbal SAT when all the items are at the same medium difficulty level. It is obvious that replacing medium difficulty items by hard items and by easy items reduces the measurement effectiveness for most of the examinees, since most of them are in the middle of the ability range.

All this suggests the following conclusion: If we really want effective measurement for both high-ability examinees and for low-ability examinees, and furthermore if the ability range in the group tested is sufficiently large, then it will be impossible to achieve our objective with any conventional test. The objective cannot be achieved simply by adding hard items at one end and easy items at the other end. It becomes necessary to try some unconventional form of testing, such as multilevel testing, two-stage testing, or tailored testing.

Before discussing such unconventional tests, consider an alternate possibility. Let us take our conventional test and score the answer sheets in the usual way. After doing this, let us divide the examinees

into three or four groups according to their scores. We can now rescore the answer sheets in each subgroup using a set of item scoring weights appropriate for that subgroup. For the highest subgroup of examinees, an appropriate scoring weight for each item will be roughly proportional to the discriminating power of the item, or to the item-test biserial correlation. For the lowest subgroup of examinees, the proper item scoring weights are quite different: the difficult items should each receive a scoring weight of approximately zero.

After rescoring each subgroup with item-scoring weights appropriate to the subgroup, the scores from different subgroups will all be put on the same scale, by conventional equating methods. Once this is done, each examinee tested will have been scored with a set of item scoring weights roughly appropriate for him. Thus each person will be measured more effectively than under conventional scoring procedures.

Although this would result in some improvement, I do not believe it is a very effective solution to the problem under discussion. If only a quarter, or a third, of the items in the test are really appropriate for low-ability students, then no amount of statistical manipulation will make this into a really good test for such students. The only way to achieve this is somehow to arrange so that such students take a full set of test items, all of which are appropriate and effective for them.

I am not necessarily urging that effective measurement of low-ability students should be a prime objective of the College Entrance Examination Board. Most of the colleges that use the College Board tests are concerned with effective measurement in the upper half or two-thirds of the score range. On the other hand, there are some colleges using these tests where most students score in the lower part of the range. Thus is may be desirable for the test to measure effectively there too. Also, it may be desirable that the test should not be a traumatic experience for those lower-level examinees who take it.

If we wish to be sure that the difficulty level of a test is matched to the ability level of the particular individual taking it, we can consider various unconventional procedures embraced by the term individualized testing. There are various names for these procedures such as computer-based testing, branched testing, sequential item testing, tailored testing, flexilevel testing, multilevel testing, and two-stage testing.

The United States Civil Service Commission is carrying out an extensive investigation into tailored testing. It has several computer terminals in its Washington office where volunteers are invited to take a tailored test. Vern Urry at the Commission tells us that this experi-

mental work is very successful. The people taking the tailored test like it better than the conventional test. Furthermore the Commission is currently able to achieve with about twenty items what formerly would require a hundred items. The Commission is making plans to use tailored testing on a nationwide basis in about five years if no unexpected obstacles are encountered.

Computer-based tailored testing is a fairly complicated procedure requiring some initial investment. There is a simple procedure called multilevel testing which is currently more readily available to all of us. An experimental study into the effectiveness of a multilevel test was recently carried out under the direction of Dr. Gary Marco, ETS. The final report on this study has not yet been issued; today I will simply describe a multilevel test.

Suppose that we have a set of fifty items all measuring roughly the same psychological trait or skill or ability. The items are arranged in five levels: a, b, c, d, e, in order of difficulty. All students start the test by answering level c. At this point they are told that if the items they have answered seemed rather difficult, they should next answer level b. If level c seemed rather easy, they should next answer level d. When they have completed a second group of items, an appropriate set of instructions is again given allowing each examinee to choose a third level of items adjacent in difficulty to the levels already answered.

Each examinee winds up taking a block of exactly 30 consecutive items (3 consecutive levels). Each answer sheet is scored in the usual fashion. There are three different possible blocks of items that an examinee may take: abc, bcd, or cde. Scores on these three blocks must be equated across blocks. This can be done by conventional methods, or by using item characteristic curve theory. Once all scores have been put on the same scale by equating, each examinee should be measured more effectively than by a conventional test, since each examinee has presumably taken items better matched in difficulty to his ability level.

It may be helpful to think of a multilevel test as if it were a three-stage test. The examinee does his own routing. This avoids the problem of scoring each stage in time to route the student to an appropriate later stage.

You can all think of various possible difficulties with such a multilevel test. Suppose an individual does not route himself appropriately. In this case, the worst that will happen is that he will be measured less accurately than otherwise. If the tests are properly equated, his expected

score will not be affected. We hope that most of the students will route themselves appropriately and thus be measured more accurately than by a 30-item conventional test.

From what I know of the results, the multilevel test tried out last fall was about as effective as expected. A detailed discussion will appear in the final report of this study, at which point the practical value of multilevel testing can be better assessed.

Another recent application of test theory in the public interest is item sampling. When examinees are sampled also, we speak of matrix sampling. Although this application is well established, many of the necessary mathematical formulas are so long and cumbersome that they have never been worked out. I would expect that the next important basic development in this area would be a computer program by means of which the computer itself will carry out the mathematics and derive the necessary formulas.

There are several other important, relatively new applications of test theory in the public interest. One of these is the design and evaluation

Figure 4.
Black (dashed) and white (solid) item response curves for item 8.

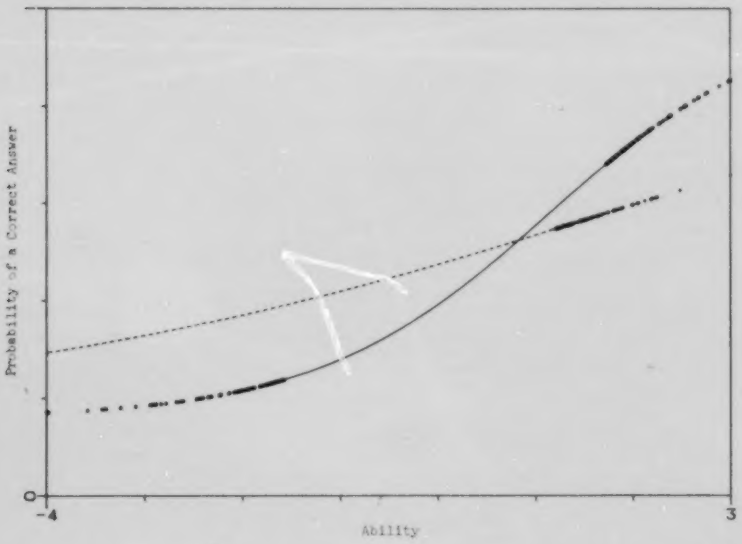
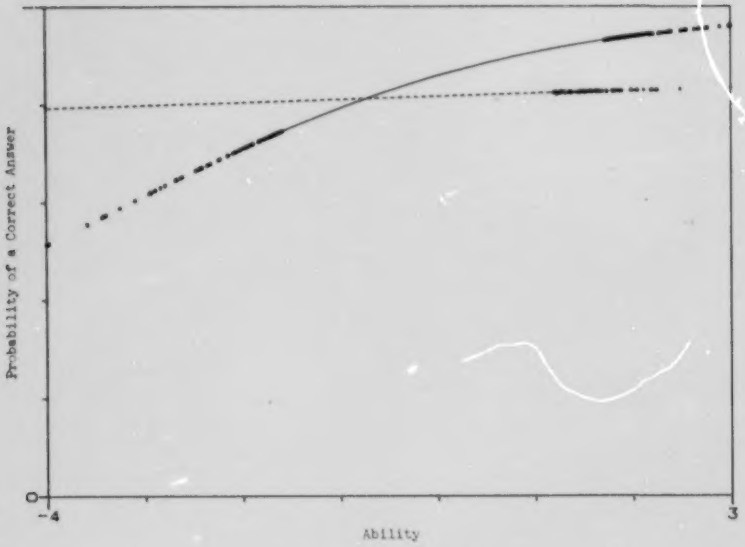


Figure 5.
Item response curves for item 2.



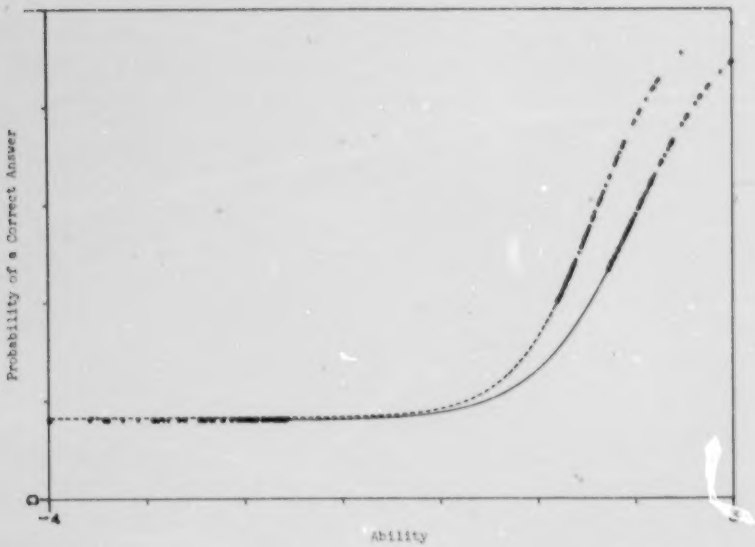
of mastery tests. My own opinion is that Allan Birnbaum's Chapter 19 in Lord and Novick[1] provides a detailed and clearly worked out theory for the design and evaluation of mastery tests. Other approaches will doubtless be effective also.

Another area, still very much in formation, is the use of tests in individualized instruction or in computer-assisted instruction. Such use of tests may come under the heading of mastery tests. I find that it is considerably different from the tailored testing discussed earlier.

In closing let me return to the question of bias, but now instead of considering test bias, let me talk about item bias. In the last three figures, the base line in each figure represents ability or skill. The curves in each figure represent the probability of success on a particular item as a function of ability level. The three figures are for three different items from the Verbal Scholastic Aptitude Test. The solid curve in each figure is for a group of white students. The dotted curve is for a group of black students.

In Figure 4 we see that high-ability white students do better on this item than high-ability black students, but that low-ability black students

Figure 6.
Item response curves for item 59.



do better on the item than low-ability white students. Black students do better than white students throughout most of the ability range.

Figure 5 shows a partially similar situation except that in this case the item is totally undiscriminating for black students. High ability black students, as determined by other items on the test, do no better on this item than low ability black students.

Figure 6 shows a difficult item on which blacks do better than whites at every ability level where there is a difference. There are, of course, other items on which whites do as well or better than blacks at each ability level.

Such items contain a bias, a somewhat complicated kind of bias. It would seem desirable to exclude such items from our tests as far as possible. Let me emphasize that the curves shown here were picked simply because they did show a definite difference between black groups and white groups. Most of the items in the Verbal SAT do not show large biases of this kind.

These curves have only recently become available as a result of a study designed by Dr. Marco. We have not yet had time to study the

test items and compare them with the statistical results. It is to be hoped that as a result of such studies, we will learn how to design items that do not show these kinds of bias.

The thread running through the various applications of test theory that I have discussed is the evaluation and design of tests for particular individuals, or for particular subgroups, or at particular ability levels. Such concerns represent worthwhile applications of test theory in the public interest.

References

1. Birnbaum, A. Classification by ability levels. In F. M. Lord and M. R. Novick, *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley, 1968. Pp. 436-452.
2. Darlington, R. B. Another look at 'cultural fairness.' *Journal of Educational Measurement*, 1971, *8*, 71-82.
3. Linn, R. L. and Werts, C. E. Considerations for studies of test bias. *Journal of Educational Measurement*, 1971, *8*, 1-4.
4. Lord, F. M. Quick estimates of the relative efficiency of two tests as a function of ability level. *Journal of Educational Measurement*, 1974, *11*, 247-254.
5. Thorndike, R. L. Concepts of culture-fairness. *Journal of Educational Measurement*, 1971, *8*, 63-70.

Testing: The Baby *and* the Bath Water Are Still With Us

ESTHER E. DIAMOND
Senior Project Director
Science Research Associates, Inc.

Having had to submit a title for my paper—"The Baby *and* the Bath Water Are Still With Us"—before I had begun to write it, I must now try to make it work...

Once upon a time there was a baby—a beautiful, smiling, unspoiled baby whom everybody admired and who, the people thought, would bring enlightenment into the world and open doors long barred to most of them. One day, when the baby was being bathed, someone noticed that the bath water hadn't been changed for a while and it had gotten cloudy and somewhat dirty. For some strange reason no one in the household was quite sure what to do about it. Some advocated throwing out the bath water. Others said the baby should be thrown out because it had contaminated the bath water. Still others argued that since both the baby and the bath water were obviously contaminated, it would be best to get rid of them both. A group of very conservative members of the household, not willing to take any risks, opted for keeping both but conceded that the dirty water could be removed a teaspoonful at a time and replaced by clean water. And so, some undetermined number of teaspoonfuls later, here we are; the baby and much of the bath water are still with us.

So much for the analogy...

We have had, over the past two decades, some enormously complex problems relating to testing. And although it is obvious that we have made some progress on a great many fronts, we cannot really say that we have taken a giant step or two forward.

Defining the Issues

The issues are quite familiar to most of us. Broadly defined, they concern the purposes of tests; the test content and what it measures; and the ways in which test results are presented, interpreted, and used.

Why do we test, particularly in the schools? In the best of all possible worlds the main purpose of measurement in the schools should be to facilitate understanding of the individual as a whole, complex, continuously developing person. Such measurement should provide information about the individual's cognitive and noncognitive characteristics, style of learning and of solving problems, and his or her needs, values, interests, and goals. Such information should also help teachers and counselors to provide the best possible instruction and guidance, and interventions designed to enhance personal development.

Unfortunately, however, this is not the best of all possible worlds, and truths, half-truths, and untruths wage a chaotic war within it. Today's tests, it is charged, do not measure the more elusive qualities of an individual, such as creativity or the ability to cope. True, but most tests—especially those given in the schools—don't purport to do so. The test title and the technical manual usually make it clear that the test is a test of reading achievement, for example, or mechanical understanding, or vocational interests. Until measures of these other qualities have been developed successfully, we shall have to be content with using, along with those test scores that are available, all other information we can gather about an individual—a highly recommended practice at all times, regardless of how much test data is available.

Another charge—in fact, probably the major charge heard against testing today—is that the test content and the resulting norms reflect the dominant culture and are insensitive to differences in experience, language, and cognitive style and the ways in which they might interact with test directions and test content. Normative data, it is further charged, make unfair comparisons that are then used to pin erroneous labels on members of minority groups, limiting their options with regard to education, career, and way of life, and perpetuating destructive stereotypes.

Few would argue that there is not one iota of truth to these charges. Tests *are* sometimes misused and their results erroneously interpreted. Individuals *have* been erroneously labeled and relegated to a very narrow set of options. Test content sometimes *does* reflect instances of bias

–both cultural and sex. Irrelevant tests *have* been used for employee selection and evaluation and for other purposes for which the tests in question were never intended.

Tests and Bias

Almost all of the charges relate in one way or another to the issue of bias–some to a greater, some to a lesser extent. To deal with the issue of bias, though, it is first necessary to know what it is we are talking about and to be sure that we are all talking about the same thing. As of now, we are far from agreement on a definition, although the literature of the past few years contains a great abundance of studies of bias and the attempts to correct it. Cleary[4] has suggested that a test is biased if scores for subgroups are consistently predicted too high or too low. *Standards for Educational and Psychological Tests*[1] alerts test users to the existence of many different definitions of bias and fairness and points out that whether a given procedure is or is not fair may depend on the definition accepted. Somewhat similar problems have arisen with regard to the definition of sex bias–both in career interest measurement (Diamond[7]; Hanson & Prediger[13]) and in achievement testing (Diamond[8]).

Breland and Ironson[3] ask: What *is* a minority? What *is* a disadvantaged applicant? The problem of classification of different minorities, they have found, is a complex and virtually insurmountable task. The DeFunis decision, for example, defined a minority as a select group of nonwhites, excluding Asian Americans except for Philippine Americans, and excluding Puerto Ricans but not Chicanos.

Ebel[9] has argued that "The bias which accounts for poor test performance by some minority persons is not in the tests so much as it is in the culture, and thus is another problem altogether" (p. 87). Even if we agree–and I don't think that test bias and cultural or societal bias are mutually exclusive–how do we go on from there? Can we afford to wait until society corrects its own biases, through a gradual process of education and change? Judging from the desegregation experience, that may be a long time–as much as one hundred years. Should we instead try interventions of various kinds–including intervention in the testing situation–wherever there is a chance that they might be effective?

Sources of Bias

If we are to do anything about bias in testing, however we define it, we should first consider its sources. Flaugher[11] has defined three principal sources.

1. *The test content.* Probably the most commonly perceived source of bias in testing is the content of the test itself: Is it biased in language? Does it lack balance in its appeal to different groups? Is it insensitive to differences in experiences or the absence of certain experiences?

2. *The atmosphere of testing.* I would enlarge this source to the society itself and place it above test content in importance. Much of the research in this area deals with the self-concept the individual brings to the testing situation and his or her perceived relationship to the larger society. Flaugher includes the amount of sophistication or experience needed to overcome idiosyncratic characteristics of the testing situation. Among these are the type of test item and the answer sheet format, which constitute the *medium* and which students must overcome in order to concentrate on the *message* of the test content itself. Other variables in this category are race (or, I might add, sex) of the examiner and perceived use to which the test results are to be put.

3. *Test use.* Biased use of test results would occur where one group is sytematically favored over the other in selection, classification, and the like on the basis of test results—whether the membership group be black, Chicano, male, female, or any other.

Although Flaugher states that women "are not the usual sort of minority group and do not have the usual sort of difficulties with testing" (p. 3), it is not difficult to see the same three sources operating with regard to sex bias. The content of the test often reflects experiences that traditional social roles have closed to women—or men—or have thoroughly discouraged them from exploring. Subtleties of the socialization process often carry over into the atmosphere of testing, where women—and, to a lesser extent, men—bring to the testing situation the self-concept that society has preordained for them. And test results have frequently been used to rule out nontraditional options and to perpetuate the status quo.

Looking for Solutions

What, then, should be done about testing? How should the controversial issues be resolved? Generally, two opposed courses are suggested:

1. Declare a moratorium on all tests and testing until the shortcomings can be eliminated.

2. Retain the tests for the information that they can provide, and at the same time encourage a program of research directed toward elimination of or control for bias and place top priority on better interpretation and use of test results.

As *Standards for Educational and Psychological Tests*[1] points out, to declare a moratorium on the use of tests requires a corresponding but unlikely moratorium on decisions—employment decisions, selection decisions by colleges and universities, and decisions based on the evaluation of various educational and social programs. But there always have been such decisions, with or without testing, and they will continue to be made. Colleges and universities, the *Standards* go on to say, will continue to select students, "some elementary pupils will still be recommended for special education, and boards of education will continue to evaluate the success of specific programs" (p. 2). The decisions, however, will be based on more subjective, less dependable methods than standardized assessment techniques. Moreover, tests that are useful for discovering abilities that might otherwise remain unidentified will no longer be available.

To assume that such decisions can be made fairly without reliable, objective measures is to assume that everyone charged with making judgments about others in our society is socially concerned, free from prejudice, and trained in the skills and pitfalls of assessment, diagnosis, and evaluation. If tests are guilty of reflecting middle-class values, will the judgments of middle-class teachers, counselors, administrators, and employers necessarily be less so? Can any of us honestly say that he or she has almost never misjudged a person's capabilities or attitudes because of some idiosyncratic mode of dress or social behavior or some unusual physical characteristic? Have our own value systems never entered into our judgments of others?

The argument in favor of a moratorium also implies that decisions are made about individuals on the basis of test scores alone. Yet test

manuals and professional articles and books on testing carry repeated warnings that tests are tools that provide objective and important information about an individual but that they do not provide all possible information and therefore should not be used alone but with all other pertinent information available.

If we adopt the second course—retaining the tests for the information they provide and at the same time embarking on a program to improve them and the ways in which they are used—what are the steps we should take? What kinds of relevant research and development are already under way?

Correcting Test Bias

Models for the correction of test bias that have appeared in the literature on testing over the past eight to ten years generally fall into one of three categories:

1. *Correcting test bias at the item construction level.* This is probably the least frequent model. It involves trying to build a bias-fair test from scratch, beginning with the instructions to item writers, before items are pretested. One example is the work of Rayman[20], who attempted to construct interest inventory items for vocationally related scales that would be balanced for response rate by sex within each scale. A similar model for achievement tests was suggested by Diamond[8].

2. *Correcting test bias at the item distribution level.* This type of model is closely related to the first type, except that it begins with the items already in hand and the item statistics for the various groups involved in the testing. Medley and Quirk[15] examined differences between black and white candidates' performance on the common examinations of the National Teacher Examinations. They constructed experimental forms and compared performance on items reflecting black culture, those reflecting modern culture, and items that were considered traditional. Differences in performance on one test made up of equal numbers of black and modern-culture items and another test consisting of traditional items only were significant for 13 of the 14 pairs of groups tested. Significant differences were also found in favor of blacks on the black-culture items and in favor of whites on the modern-culture items.

   Echternacht[10] compared the distributions of transformed p-value

differences for independent pairs of groups with a hypothetical normal distribution, using the obtained mean and variance of the differences as parameters. He considered the test biased if points on the actual distribution fell outside the bans around the hypothetical line whose width is determined by sample size and significance level.

Angoff[2] describes several studies, including his own, in which bivariate plots of transformed p-values were examined for item x group interaction. Angoff also mentions the possibility of building a test on the basis of a common core of items "broadly relevant to the educational objectives of society generally and the individuals for whom it is intended" plus items specific to the curriculum of each of the component groups but not the group as a whole (p. 26). With such balance, Angoff maintains, no one group would have an advantage across the total test.

In one study described by Angoff, involving black and white groups, item x group interaction for inter-race scatter plots decreased when groups were matched on an external variable. This result suggests the possibility of matching groups on socioeconomic status, expressed as a composite of parental occupational and educational levels. Angoff warns, however, that the designations for these levels might not have exactly the same meanings for blacks as for whites.

3. *Statistical models for the correction of bias.* Various statistical models for dealing with test bias have been proposed—by Cleary[4], Cole[5], Darlington[6], McNemar[14], Thorndike[22], and others too numerous to mention here. The entire Spring 1976 issue of *Journal of Educational Measurement* was a special issue, *On Bias in Selection.* In that issue the Novick and Lindley utility model is described by Novick and Petersen[17]. Cleary's model was referred to briefly earlier in this paper. Cole's model suggests that if both a member of the majority group and a member of the minority group could succeed if selected, any procedure is unfair that does not present each with the same probability of being selected. It requires that different predictor cut-off points be chosen for each group. Darlington's model employs a single correction factor whose variable weight, determined by a set of factors important to the selecting institution, would be added to the criterion scores of the lower-scoring group. McNemar's model employs a regression equation based on the groups combined, with group membership included as a predictor. Thorndike suggested that the percentage of an applicant group to be selected be the same

as the empirically determined base-rate of success for that group. These models are probably part of the necessary groundwork for a temporary solution of the problem of bias within the present framework of inequality of opportunity. Many of these models, however, are in conflict with each other in one or more respects, and it may be a long time before one is developed that wins the widespread acceptance needed to put it into general practice.

Something should be said here, too, about the various attempts over the years to build "culture-free," "culture-specific," and "culture-fair" tests. These usually refer to so-called tests of intelligence rather than to tests of achievement, but are sometimes suggested as replacements for standardized achievement tests. I think that there is general agreement that it is virtually impossible to build a culture-free test; no group lives in a cultural vacuum. Culture-fair tests might fit some of the models for correcting bias at the item construction or the item distribution level. Nonverbal culture-fair tests, as Ornstein[18] points out, generally fail to reflect the full range of a child's mental abilities. Moreover, the child who has trouble with verbal tasks generally has trouble dealing with such perceptual tasks as classification, selection, and arrangement. As for the culture-specific Black Intelligence Test of Cultural Homogeneity (BITCH), it has been criticized by Ornstein and others as measuring a very limited amount of special information useful for functioning in the ghetto. The ability to label, categorize, conceptualize, and solve problems—an ability important for *all* children if they are to succeed in school—is not dealt with.

Another problem that further complicates the already complex task of constructing a model for correction of bias or building a test controlled for bias is the fact that there are in the United States a great many minority cultures, some of which account for only a fraction of one percent of the population. Even among the larger cultural minorities there are differences within groups. The Spanish-speaking child of Puerto Rican parents, for example, is different from the Spanish-speaking child just this side of the Mexican border. There are comparable differences between the various Asian groups. If we try to assign everyone to a clearly defined group, there will be too many groups, most of them with relatively small numbers, to yield any meaningful analyses. If we establish only a few major groups, we may not improve the situation very much. Moreover, there appears to be considerable evidence that the differences be-

tween socioeconomic status groups within a culture are much larger than the differences between cultural groups as a whole.

Improving the Tests Themselves

While we cannot hope to fully eradicate systematic inequities in test performance until inequities in opportunity have been eradicated, there are many ways in which tests can be and are being improved.

1. A number of publishers have undertaken a reexamination of items in existing tests, with the assistance of qualified black and other minority group reviewers. Items with obvious language or content bias are being edited or replaced wherever possible, and specifications for items for new forms or new tests are being written with concern for possible bias. Tests are also being reviewed for sex bias.

2. Biographical data and other self-reported descriptive information are being used increasingly in combination with cognitive measurement for self-assessment and future planning as well as for improved prediction.

3. Work on adaptive testing, tailored to individual ability level and other characteristics, is making progress.

4. Advances in computer capabilities have made possible comparable advances in testing techniques such as branching and the provision of immediate feedback from the computer.

5. Criterion-referenced tests enable us to determine to what degree an individual has mastered a particular skill or content area rather than how that individual compares with others, thus eliminating the kinds of objections that are made to norm-referenced testing. Ironically but understandably, however, some publishers of criterion-referenced tests are being asked to supply norms as a kind of reference point for the interpretation of the criterion-referenced scores. Such normative data should be acceptable to all concerned if it involves group rather than individual comparison. Schools want to know whether a given average score indicates strength or weakness in the domain measured, and group norms give them a picture

of *relative* strengths and weaknesses. The danger, however, as Popham[18] points out, is that users of criterion-referenced tests will rely on normative data as a *determiner* of performance standards.

6. Progress has also been made in diagnostic testing and evaluation. Expanded computer capabilities have made possible detailed and highly sophisticated item analysis for local and special groups and for individual students. Growth curves in specific skills can be drawn by the computer. The effects of various kinds of interventions can be analyzed along a number of dimensions.

7. There has been a growing trend toward the use of tests for placement and classification, as opposed to selection, and a growing emphasis on decision-making skills that will help individuals use data from tests and other sources to make for themselves many of the decisions that have traditionally been the responsibility of the school, the employer, or other institutions.

These developments are encouraging, but there are still unfulfilled needs to be met. Some have been described by Gordon[12], Mercer[16], and others. We need measures that will provide information about a much wider range of abilities and characteristics than present measurement provides—measures of vocational, social, and interpersonal competencies; of creativity, which we have not so far even defined successfully; of cognitive style, or how the individual processes information and generates responses. We need to know how best to weigh all the information we have about an individual in order to enable him or her to make the best possible decisions. We need to find ways to solve the dilemma posed by prediction based on the past that works to perpetuate the status quo. We need item analysis programs that enable us to look at the incorrect choices children mark on tests to see whether an individual or group pattern emerges that might be of diagnostic significance. These are only some of the needs. The list is virtually endless.

Improving the Use of Tests

No matter how much we improve the quality and sensitivity of our tests we will have gained little if the way in which they are selected and used is not also improved. This must be a joint responsibility of both test

publishers and the institutions using the tests, with the publisher providing the interpretive material, descriptive information about the test and its purposes, and suggestions for use, and the institutions providing the necessary training in test use, possibly with help from the publisher.

A school testing program, for example, should be based on the joint decisions of those who will have to implement it and interpret the results. This means involving counselors and teachers, or at least representatives from among them, in addition to the school principal or the superintendent of schools and anyone else who will play a major role in the program.

Questions to be discussed by these individuals include:

1. What is the purpose of the testing program? What is it the school needs to know, and which tests can help supply the answers?
2. Do the tests under consideration fit the intended purposes of the program? That is, do the tests measure the traits or content areas or attitudes that the school wants to know about? Technical manuals and interpretive information should provide answers to this question.
3. Does the content of subject-matter tests—whether norm-referenced or criterion-referenced—match, in general, what students have been exposed to in their course work?
4. Is the reading level such that most students can be expected to understand the language of the test?
5. Are the directions to the students clear so that the average student will not have difficulty following them?
6. Can the results be used for diagnosis of specific difficulties as well as for general measures of achievement, ability, and so on?
7. Are the hidden biases—overall content slanted to white middle-class values and culture, or to traditional sex-role behavior?
8. For standardized tests, are the norms provided generally useful for the particular school population? If not, are local or other appropriate norms available, or is information provided that will suggest how to interpret the results for the students?
9. How does the school plan to use the test results to help students? How will the results be communicated to students and their parents?

10. Do the tests meet the essential requirements of the APA *Standards for Psychological Tests and Manuals?* What does Buros' *Mental Measurements Yearbook* say about them?

Parents and students should also be told something about the testing program and why it is being given. Some publishers have prepared letters to parents for this purpose. If these are not available, the school should prepare its own. If student information booklets containing a description of the test and sample items are available, they can be used in a brief test orientation session with students, to put them at greater ease in the testing situation. Filling in sample answer sheet grids well ahead of the testing date also helps reduce irrelevant sources of error on the test itself.

When test results are available, all who will be involved in the interpretation should be briefed on the results and what they mean. Report forms, profiles, bands of confidence, the meaning of percentiles, the differences between measures of ability or achievement and measures of interest—all these should be understood by teachers and counselors before the results are disseminated. The school might also want to consider involving parents and students, especially students at the high school level, at some point. Parents will want to know what the results mean for the child. What new information has the test added to what is already known about the child? Are there contradictions between the test results and other information? If so, how can they be explained? Finally, both parents and students will need reassurance that test results will be used constructively—that a low score on a reading test means, usually, only that the child needs help with reading.

Conclusion

I hope I have succeeded in demonstrating that, although the baby *and* the bath water are still with us, the bath water is much cleaner now than it has been for a long time. And a lot of effort is going into making it still cleaner.

I'd like to close with a quote from Theodore Sizer's[21] conclusion at the ETS Conference on Testing Problems six years ago:

"...the testing fraternity needs to concentrate on the effects of class, race, and ethnicity on the development of skills and attitudes. It needs to help us understand how these factors influence human development

over time. It needs to suggest ways of lessening those influences that narrow a youngster's options, and ways of measuring the child's progress in increasing his options.

"Testing must not in a benign way serve as a device to preserve the social status quo. On the contrary, it must be used to illumine current social rigidities—and to help us finally break out of them."

References

1. American Psychological Association. *Standards for educational and psychological tests.* Washington, D.C.: APA, 1974.
2. Angoff, W. H. The investigation of test bias in the absence of an outside criterion. Paper presented at NIE conference on test bias, December 1975.
3. Breland, H. M. & Ironson, G. H. DeFunis reconsidered: A comparative analysis of alternative admissions strategies. *Journal of Educational Measurement*, 1974, *13*, 89-99.
4. Cleary, T. S. Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement*, 1968, *5*, 115-124.
5. Cole, N. S. Bias in selection. *Journal of Educational Measurement*, 1973, *10*, 237-255.
6. Darlington, R. B. Another look at cultural fairness. *Journal of Educational Measurement*, 1971, *8*, 71-82.
7. Diamond, E. E. (Ed.) *Issues of sex bias and sex fairness in career interest measurement.* Washington, D.C.: U.S. Government Printing Office, 1975.
8. Diamond, E. E. Minimizing sex bias in testing. *Measurement and Evaluation in Guidance*, 1976, *9*, 28-34.
9. Ebel, R. L. Educational tests: valid? biased? useful? *Phi Delta Kappan*, 1975, *57* (2), 83-89.
10. Echternacht, G. A quick method for determining test bias. *Educational and Psychological Measurement*, 1974, *34*, 271-280.
11. Flaugher, R. L. *Bias in testing: A review and discussion.* TM Report 36, ERIC Clearinghouse on Tests, Measurement, & Evaluation. Princeton, N.J.: Educational Testing Service, 1974.

12. Gordon, E. W. Affective response tendencies and self-understanding. In *Proceedings of the 1973 Invitational Conference on Testing Problems.* Princeton, N.J.: Educational Testing Service, 1974.

13. Hanson, G. R. & Prediger, D. J. The distinction between sex restrictiveness and sex bias in interest inventories. *Measurement and Evaluation in Guidance*, 1974, *7*, 96-104.

14. McNemar, Q. On so-called test bias. *American Psychologist*, 1975, *8*, 848-851.

15. Medley, D. M. & Quirk, T. J. The application of a factorial design to the study of cultural bias in general culture items on the National Teacher Examinations. *Journal of Educational Measurement*, 1974, *11*, 235-245.

16. Mercer, J. R. I. Q.: The lethal label. *Psychology Today*, September 1972, 44-47.

17. Novick, M. R. & Petersen, N. S. Towards equalizing educational and employment opportunity. *Journal of Educational Measurement*, 1976, *13*, 77-88.

18. Ornstein, A. IQ tests and the culture issue. *Phi Delta Kappan*, 1976, *57* (6), 403-404.

19. Popham, W. J. Normative data for criterion-referenced tests? *Phi Delta Kappan*, 1976, *57* (9), 593-594.

20. Rayman, J. R. Sex and the single interest inventory: The empirical validation of sex-balanced interest inventory items. *Journal of Counseling Psychology*, 1976, *23*, 239-246.

21. Sizer, T. R. Testing: Americans' comfortable panacea. In *Proceedings of the 1970 Invitational Conference on Testing Problems.* Princeton, N.J.: Educational Testing Service, 1971, p. 21.

22. Thorndike, R. L. Concepts of culture fairness. *Journal of Educational Measurement*, 1971, *8*, 63-70.

Luncheon Address

One Man's View of Testing

WILLIAM RASPBERRY
Columnist
The Washington Post

If I occasionally find myself rebutting attacks on standardized tests, it is not because I think the tests are that great. It is because I think they are often attacked for the wrong reasons.

I am thinking, for example, of the attacks premised on the fact that blacks and other disadvantaged minorities do less well on standardized tests than do middle-class white children.

I am thinking of the blitz of the *National Elementary Principal* magazine [Vol. 54, No. 6; July-August, 1975] which, in a single issue, devoted 18 articles and an editorial to the subject of standardized testing and managed to find not one single good thing to say about it.

I am thinking of the assaults by people who have a vested interest in my not finding out how well, or how poorly, the schools are doing in their primary job of educating children.

I am thinking of people who scream cultural bias without the faintest idea of what they mean.

I am thinking of people whose objection is to policies, but whose attack is on tests designed to effectuate those policies. They denounce screening of fully qualified applicants to graduate school, for instance, simply because there are fewer spaces than applicants.

And so, although I happen to believe that the test makers are not doing nearly a good enough job of devising tests or helping those who administer them to understand their proper use, I frequently find myself opposing those who attack testing.

I found myself in verbal combat with the former superintendent of schools in Washington, D.C., Mrs. Barbara Sizemore, when, after recently published scores showed that our children were performing poorly, she proposed an end to testing.

I pointed out that there might have been all sorts of reasons why it

would be unfair to compare test scores of children in Washington slums with those of children in Palo Alto. But it did seem to me, I said, that some other explanation was called for when the results showed that Washington children were doing less well in reading and math than Washington children had done the year before and the year before that. When tests reveal trends, I said, it seems they are trying to tell us something. As a rule, I count it better to listen than to throw the tests away.

True, there were problems with the tests. As Mrs. Sizemore pointed out, there is no assurance that you are testing the same children from one year to the next. Nor, without some attempt to chart migration patterns and changes in the socioeconomic patterns of the student population, can one assume that test results reflect what happens in the schools.

But whatever is wrong with the tests, there are some things they can do. They can tell, within limits, how the children in your hometown stack up scholastically with the children across town or across the country. And they can tell you how the children in your schools stack up with their predecessors in those same schools, or what happens to a particular class of students during its school career.

These are things worth knowing. But some people do not want us to know them. That was my suspicion when I read the *Elementary Principal* magazine I mentioned earlier. Standardized tests, a dozen and a half authors concluded, "destroy" children. The tests, they said, are illogical, misleading, and may inspire cheating—comparing people to one another along a single scale of ability is fundamentally demeaning and unfair.

Not only do the tests do badly what they allege to do; what they allege to do should not be done in the first place. "...standardized science achievement tests for the elementary school are almost uniformly poor in quality. They are incorrect, misleading, skewed in emphasis, and irrelevant....

"The scores purport to be measures of the educational health of a community or a school. But, in fact, it would make as much sense to take the blood pressure of each student, apply the usual statistical procedures, and publish the results district by district, to measure the health of the student body."

The articles took the usual potshots at individual test items, many of which are incredibly bad, and often concluded that any test containing such items is worse than useless. For instance, this multiple-choice item:

Many kinds of plants are not able to live in the desert because of the

high temperature
low rainfall
bright sunlight
poor soil.

Any scorer who marked any one of those test choices wrong ought to be fired. But some of the attackers took exception to virtually all multiple-choice questions, including this one:

What do scientists use to make small things appear larger?

a barometer
litmus paper
a balance
a microscope.

I thought it was a fairly unambiguous item. But the author who cited it had this comment: "What are 'small things': small differences in pressure? small changes in acidity? small weights?"

In other words, according to the author, this is yet another ambiguous item. Not to me. If I gave the question to a science student and asked him to come up with a justification for each possible answer, that would be one thing. But if I gave him the question, and made him understand that there was only one acceptable response, and that he would have no opportunity later on for justifying exotic answers, that would be another thing altogether. In that case, if I asked him what instrument made small things appear larger, and he said "barometer," I would think he either was not very bright or that he was being a smart-aleck.

One point is made again and again. Norm-referenced standardized tests do not tell you what to do about low achievement, nor do they prescribe remedies. My response is that the instrument panel on the dash board of my car includes a speedometer, which does not tell me why my car is not going faster; a temperature gauge, which does not tell me why it is running hot; and a clock, which does not tell me why I am late and what alternate routes I might take to make up the time. But it does not follow that these are useless instruments. Sometimes it is very helpful to know something is wrong.

Now, what of cultural bias? Well, it depends on what you are talking about.

One Man's View of Testing

Everybody knows what cultural bias does: it causes inner city blacks and other disadvantaged children to score poorly on intelligence, aptitude, and achievement tests. But hardly anybody seems to know exactly what it is, or whether it is a correctable condition. And as a result, there is a growing demand for its solution by radical surgery: get rid of the tests. Unfortunately, those who tend to be most victimized by cultural bias, *whatever* it is, are least in a position to dictate an end to testing.

There are two main propositions concerning testing and cultural bias. They often coexist in the same argument and occasionally are commingled in the same sentence.

The first is: standardized tests, because of cultural bias, do not accurately measure the capabilities of black and other minority test-takers. The second is: standardized tests may be a more or less accurate way of testing capabilities (though not native intelligence) but, because of cultural bias, they test those capabilities at which the middle-class and white, rather than the poor and black, tend to excel.

The first says test do not do very well what they allege to do, as far as blacks and minorities are concerned. The second says the tests are designed to uncover virtues which the dominant society deems important and not others which it considers less important. One of the things that is rated important is the degree to which applicants have absorbed and internalized the dominant culture, including those things normally taught in schools.

When you put it that way, it becomes clear that the test is supposed to be culturally biased. That is one of its purposes. It might not tell you whether the learning, the acculturation, took place in school or at home. It will not measure those aptitudes and achievements that the test-designer was not looking for, and it will not tell you anything definitive about native ability.

But if your purpose is to know how much of "A" a child has absorbed in order to know when to proceed to teach "B", standardized tests can be useful—unless, of course, proposition one is true, in which case the test will tend to undermeasure the knowledge of black children.

Because of the confusion over what cultural bias is, attempts to remedy it have shot off in a number of directions. Some have attacked tests that purport to measure reasoning ability as being, in reality, tests of socioeconomic status and vocabulary. Take, for instance: Candelabra is to candle as chandelier is to (a) book, (b) Ben Hur, (c) light bulb, (d) elaborate. Some critics would see this as an obvious example of cultural bias. What does a kid from the ghettos or the barrios know about chan-

deliers and candelabra? They might recast the question to read: Finger is to wrist as toe is to (a) elbow, (b) foot, (c) tap dance, (d) ankle. And they might be stunned to see their ghetto youngsters miss that one too.

For it is beginning to appear, to me at least, that the cultural bias in tests is not always in the vocabulary and content but in the form. The bias may be in the test question as a device for uncovering reasoning ability.

Robert Williams' BITCH test (Black Intelligence Test of Cultural Homogeneity) sidesteps the problem by testing for vocabulary only. And since the vocabulary is based almost exclusively on ghetto usages, Dr. Williams' test also produces higher scores for blacks than for whites—a sort of reverse cultural bias.

But not really. Vocabulary testing is too limited a solution; it does not tell us enough. Dr. Williams says he is working on tests that will do for questions in logic what the BITCH test has done for questions in vocabulary. He did not say what he will call this second-generation test, and I did not ask.

No matter. The solution is not to come up with cute things that reverse the usual black-white scoring patterns. The solution is to do what we can to give poor black and other minority children the sort of background and support knowledge that have currency in the country, to increase their opportunities for escaping the crippling effects of poverty, and to help them pass tests.

Meanwhile, there are a few things I'd like to talk to test makers about. I would like to hear them explain the necessity of distributing populations. My understanding is that one of the requirements of standardized tests is that they distribute the tested groups into bell-shaped-curves. They are very clever at doing that—at making each individual test item do that. But I am not sure I understand the point of it.

What would seem to me to make more sense is to devise tests calculated to determine how much of what is to be taught has in fact already been learned. That way, you would not have to throw out an item just because too many people got it right. You would have a device that tested children against the course material, rather than against each other. Comparisons would still be possible, of course, but that would not be the whole point.

It strikes me as particularly pointless to construct tests—those bell-shaped monsters—for graduate record exams, medical aptitude exams and LSATs, because much of the weeding-out process already has been accomplished.

I would like to see the test makers and test users agree to try to come up with an instrument that is capable of establishing a cut-off point below which success would not be predicted but which would make no effort to rank those who score above the cut-off.

And I would like to see the test makers do a much better job than they have done in increasing public understanding as to just what their tests are supposed to do.

Afternoon Session

The Student and Testing

THELMA T. DALEY
Career Education Specialist
Board of Education of Baltimore County, Maryland

As an organizational person, I have been in plenary sessions when the quiet and seemingly perfunctory motions—sometimes bordering on being soporific—have erupted as forcefully as a supposedly sleeping volcano suddenly found belching and emitting tons of "frightening" lava. Likened unto the volcanic action has been the move to place a moratorium—a five-year moratorium, a one-year moratorium, an indefinite moratorium—on all tests—standardized tests, that is. I have witnessed the widespread debate on the various issues concerning testing. I have heard columnists, commentators, journalists, experts, and neo-experts on the subject.

The great debate continues, and although the voice of the student may not be seen or headlined as one of the great debaters, the issues are irrelevant unless they relate to the human test taker—the student. In fact, I wonder why the widespread debates seldom see students as debaters; a search of records does not reveal a moratorium called by students.

In an educational era of accountability, students may read about their achievement (or lack of achievement) in the major newspapers almost on a daily basis or may hear their collective performance discussed over the local television channels. A typical example is the front page story in the Tuesday, October 19, 1976 edition of the *News American* (Baltimore), "Students Still Lag in Tests," which in part states that "pupils' scores on standardized tests of basic reading and math skills showed some improvement last year, but average scores for three of four grades tested remained in the bottom 30 percent of a national sampling."

Tests are designed, manufactured, and distributed for takers. Not all takers of tests are students, nor do all students necessarily take tests. However, it has been stated that, as a nation, we administer over 200 million achievement tests each year. This figure represents only about 65 percent of all educational psychological testing that is carried out.*

In earlier treatises, principals reported that some standardized tests were given in their schools each year.[6] In 1961, Goslin reported that 100 million ability tests a year were being taken by persons in educational institutions.[7] Later, in 1964 it was estimated that 150 million to 250 million tests a year were being administered.[1] Of 714 elementary school principals in the *Russell Sage Report*, only one reported that his school had not had plans to initiate a standardized testing program.[9]

In addition, the Coleman survey[2] reported that over 90 percent of the nation's pupils were in schools where intelligence and achievement tests were given at both the elementary and secondary levels.

Besides intensive testing programs, external testing programs, such as the Preliminary Scholastic Aptitude Test (PSAT), the National Merit Scholarship Qualifying Test (NMSQT), the Admissions Testing Program (ATP) of the College Entrance Examination Board (CEEB) and the American College Testing Program (ACT), Armed Services Vocational Aptitude Battery (ASVAB), the General Aptitude Test Battery (GATB), Civil Service Examination, Betty Crocker Search for Leadership and Family Living, all add to the number of tests administered in the school each year. This does not take into account the tests given at midterm, at the end of a unit, or the test given vindictively as a disciplinary measure.

With the increasing number of tests and the growing quest for the raison d'etre by students, one must have available the *why* for the test and the proposed use of the results. Typical uses (though many times given in a circuitous, incomprehensible way) may be (1) to select for college admission, (2) to group, (3) to identify needs, (4) to help students select courses, (5) to aid in career planning, (6) to evaluate programs, and (7) to provide information which might be helpful in securing facilities, gaining new resources, and providing research data.

The student cares very little about the research data, the accountability studies, the evaluation of programs. The student does care if he/she can visualize immediate, concrete, relevant uses.

I have witnessed large testing sessions in school auditoriums with lap boards serving as improvised desks, very poorly defined test goals (other than that the test was required of all tenth graders and it could be used to predict the next levels of achievement), and students who could not care less. Students quickly exhibited their displeasure nonverbally by rapidly running through items timed for 20 minutes in less than 10 minutes and spending the remainder of the time buckling lap boards, while the top 2 percent studiously raced against the ticking stopwatch,

and proctors silently but forcefully moved from row to row with bucklings commencing as fast as a buckler was silenced. Students today do not view massive general achievement or general scholastic aptitude testing as relevant to their needs. They will quickly ask, "What good is that to me?"

However, in an era of accountability, legislatively mandated, systems that had all but abandoned statewide or unit-wide testing programs have once again enacted them. In my own state's accountability program, the implementation plan required the establishment of a comprehensive and uniform statewide testing program. The Iowa Test of Basic Skills (ITBS) and the Cognitive Abilities Test (CAT) were selected as the statewide assessment instruments. Since the spring of 1974, all pupils in grades 3, 5, 7, and 9 have been tested on the ITBS and the Nonverbal battery of CAT; and the Maryland Basic Skills Reading Mastery Test has been assigned grades 7 and 11 as of the fall of 1975-76.

It is hoped that, as an important aspect of this assessment fabric, the results will provide teachers and schools with a basis for improving the quality of their efforts on behalf of the students. However, the capacity of a system to generate data is usually greater than the capacity of teachers to use the data.

Let me advance to some nontechnical aspects of testing that very much affect the student.

The Administrator—The Interpreter

The person who administers the test may have a negative effect on the examinees or the students. Sacks[14] found that students' scores increased if a good examinee-examiner relationship was established prior to a test.

Some writers, such as Padilla and Gazda, allude that the examiner can maximize or minimize the child's performance (on an individual test) by his or her actions. Similarly, by misinterpreting the child's responses, the examiner can significantly raise or lower the final individual intelligence (IQ) score.

In mass testing, such as accountability testing, many times teachers are examiners who have never been involved before and who may not have gone through a full orientation. Lack of knowledge and general information on the part of the examiner is ultimately detrimental to the student. Although I have no data to prove it, many teachers approach

testing sessions with very negative attitudes. In fact, when the notice arrives that a SCAT Test will be given by all English teachers, many are heard to exclaim, "Those things! There goes my good English period." This attitude is bound to be reflected in the students' perception.

Somewhere along the line, test administrators and test interpreters must meet minimum standards. This is recommended for standardized tests; however, I would go a step further and recommend that all teachers be inserviced in test making, test taking, test administration, and test interpretation.

In the McCarthy[10] study, third and fourth grade children who wrote a composition on "The Best Thing That Ever Happened To Me," prior to a test, averaged four to five points higher than their scores on the same test taken after writing on "The Worst Thing That Ever Happened To Me." Tyler[16] showed that an examinee's experience immediately preceding a test affected his/her test performance. Kirkland[9] stated that a "warm" versus a "cold" interpersonal relation, or a rigid and aloof relation versus a natural manner on the part of the examiner, may affect the examinee's responses. So, in fairness to the student, the examiner-interpreter approach must be addressed.

The Student and The Purpose

We test for many, many reasons; however, the student deserves to know why the test. ACT and SAT are popular because the purpose is clearly defined and understandable (not necessarily acceptable) to students. The PSAT/MSQT purposes are understood but become a major disappointment to students when the financial aspects peter out for the majority. Many are led to take the test with the hope that scholarships might be at the end of the rainbow, only to find out that the rainbow never appears.

There is considerable anxiety and tension associated with the taking of tests. In my counseling experiences, I witnessed students who have literally become ill on test days. Some of these same students faint and become hysterical on report-card days.

There are many hidden reasons why tests are given. Sometimes the scores are used to rule on eligibility for a basketball team; another time they might mean meeting graduation or grade-level requirements; they may mean entry into the Armed Forces, a job, or acceptance at the college of one's choice; or they might mean remedial or prescriptive

work based on the diagnosis. Whatever the purpose, the student should be fully apprised. If the test is to satisfy parental pressures, this too should be clearly displayed.

If a test is given for diagnosis, the student should see the results via a developmental program. As an example, in Maryland's accountability program, reading test results are being used to select schools to receive special assistance. This year a special project, called Project STAR (Standards Technical Assistance Resources), is in operation with 11 elementary schools showing a need in the reading area. Four specialists in the areas of reading, language development, guidance, and community involvement, plus a STAR resource teacher in each of the schools, are working with the local staff to assess their current reading programs and develop a plan for upgrading student proficiency to state standards. Integral to the project is a monitoring and evaluation system to measure growth of student achievement and staff development.

Neulinger[13]—in looking at attitudes of American secondary school students toward the use of tests—found that anti-test sentiment is neither ubiquitous nor consistent. His data showed that not every student, or every group of students to whom we administer a test, holds negative opinions about testing. His findings did indicate, however, that a student is quite likely to be inconsistent in his or her attitude toward testing. One may favor testing in one context and disapprove of it in another. Neulinger found that students' attitudes toward testing were related to social background and personality characteristics. He interpreted his findings to indicate that a student who is a member of the lower class, from a less well-educated background, who is less bright and knows it, who has limited aspirations and views of the world in fatalistic terms, reacts to tests quite differently from the respondent who is from a better educated background, who is bright and knows it, has set high goals, and thinks the world will conform to his or her wishes. For the upper class respondent, tests helped him or her to identify as a member of the elite. Tests were instrumental in getting the student into the better schools.

The student in the lower socioeconomic and less educated domain saw the test as identifying him or her—but not as a member of the elite. The identification was the equivalent of being degraded. The school, which is supposed to upgrade his or her abilities (as students see it), condemns the student before he or she gets a chance. The test excludes him or her from places of higher learning.

Neulinger concluded that students saw tests as being used by society

as a tool to differentiate among people in ways that have real consequences. Only to the degree that society is fair and just in making these discriminations will people agree that it is fair and just to use tests.

Kirkland[9], in a study on the effects of tests on students, stressed that the student was the one whose status in school and society is determined by test scores and the one whose self-image, motivation, and aspirations are influenced. Tests do affect the self-concept of students, but it is important to note that the way a person views himself/herself also influences test behavior.

In terms of motivation, exactly how students are motivated by tests has not yet been conclusively demonstrated. Most findings indicate that feedback from tests promotes learning, assuming that the student attempts to do well on the test. Students with negative scores detest the frequent feedback which tends to increase the level of low motivation.

Level of aspiration seems highly related to self-concept and motivation. Moss and Kagan[11] in their longitudinal study of intellectual progress and achievement, concluded that the child who attains scholastic honors is rewarded by those around him and that this experience frequently leads to an expectancy of future success for similar behavior, thus increasing the probability that the child will continue in such tasks. Failure would result in the opposite behavior—such as avoidance or withdrawal.

As individuals meet with success, their goals and aspirations rise in accordance with their increased confidence. Students who were tested most often and best informed about their performance were the ones most motivated to acquire additional information.

Anxiety is another big issue with students. There is considerable tension and anxiety about taking tests. Findings have indicated that anxiety scores correlate negatively with I.Q. and achievement for the so-called middle and low I.Q. groups.

The Student and the Testing Environment

Most tests are given in such ungodly places as the cafeteria, with hinged backless bench seats amid the rattling of huge metal vats and the aromatic odors of near-done meats, cooling desserts, and boiling soups—the day's menu. Many are given in dimly lit auditoriums, and occasionally the gym is readied with chairs and proctors. Long time limits and the absence of independent divisions within the test sometimes

make it impossible to administer it in the available site. Although at this writing the College Board's Blue Ribbon Panel has not aired its reasons for the score decline, I feel almost assured that testing environments may play a role.

The Student and the Score

In a recent report, Roy Forbes, Director of the National Assessment of Educational Progress, has pointed out that no testing instrument reveals everything about the quality of education students are receiving. I can recall the nervousness, the excitement, the hugs, the tears, the absolute look of failure when students have received test scores. In a quote by a student cited by Cottle[3], a young man who had just learned of his performance on a set of standardized achievement tests said, "If you eliminated money in our society, you could eliminate tests and all the test scores." He said, "So, to be American means that you have a lot of money. No matter what you earn, you aren't satisfied until you have more than the next guy. That's the same thing with tests. Giving us our score isn't enough, they have to give us the percentile rank as well. Nobody's supposed to get 690 and think they're really special. The counselor tells them right away that 690 may sound good, but it's *only* the 80th percentile. You have got to have money and you have got to have IQ, PSAT, and SAT points. Americans love numbers and quantities. Big is the name of the game. Produce and get bigger. Inches, pounds, dollars, points on tests, are all anybody cares about, even the minority students in our school. Nobody asks whether they are happy. All people want to know is whether their achievement scores have gone up, or how many points they scored in a basketball game."

The social consequences of the score has become a new area of interest in this decade. The issues, according to Ebel[4], center around such social consequences of testing as:

1. They may place an indelible stamp of inferior intellectual status on a child, ruin his/her self-esteem and education motivation, and determine his/her social status as an adult.
2. They may foster a narrow conception of ability and reduce the diversity of talent available to schools and society.
3. They may place education and the destinies of individual human beings under the control of test makers.

4. They may encourage impersonal, inflexible, mechanistic processes of evaluation and determination.

For the student, in too many cases, there is finality in the test score. The student is marked, grouped, or tagged. The score is placed on the cumulative record card, and teacher after teacher reaffirms his/her belief in the student's ineptness; later, employers see the score and readily accept that a retardate is applying, and parents get the cold score and quietly brood, wondering what they did wrong in prenatal care and subsequently develop guilt feelings. Consequently, they overindulge the child and, ultimately, foster negative behavior. One little score goes a long, long way.

Folds and Gazda[3] found that individual test interpretation, small group test interpretation, or written test interpretation resulted in more accurate self-estimates of test scores than was found in control groups receiving no information. I contend that accompanying every test must be a descriptive supplement dealing with creative, informative, and positive ways of revealing test scores to parents and students, and also to teachers who quite often forget the interpretation. Descriptive transparencies, demographics, and clear language are desirable tools that counselors and teachers welcome along with test results.

The State of Maryland has developed an occasional paper on accountability entitled, *Improving Student Attitudes and Skills for Taking Tests.*[15] Among others, the publication stresses that teachers, even the directors, should know the characteristics of the students, create a supporting environment, and avoid interruptions. Teachers are encouraged to prepare students for taking tests, tell them why they are taking the tests, how the results will be used, and how the tests are scored. They are urged to train students how to take tests, to teach them the specific thinking skills required on tests, and to inform them of the teacher's role during the test. They are urged to simulate test-taking conditions.

I remember very vividly a special education student whose name was Cy. Cy was a talented, tall, black, restless male who played the guitar, the drums, and sang. Cy hated, literally hated, his special education classes—but tests said that was where he belonged. Cy defied the scores and roamed the halls, devised ways of avoiding the teachers on hall duty, slipped to the shop to create, slipped to the music room to syncopate, slipped to the art room to watercolor, and slipped to the gym to make two points. Cy finally slipped out of sight because his tests labeled him special—labeled him dumb.

In the words of the NAACP *Report on Minority Testing*[12] tests must predict accurately what they promise; tests must measure adequately the content of the area they purport to cover; and the testing program must be capable of leading to prescriptions which result in positive growth for the persons (the students) being tested.

These are my reflections on testing and the student.

References

1. Brim, O. G., Goslin, D. A., Glass, D. A., Goldberg, I. The use of standardized ability tests in the American secondary schools and their impact on students, teachers, and administrators. *Technical Report 3*. New York: Russell Sage Foundation, 1969.
2. Coleman, J. S. Equality of educational opportunity. Washington, D.C.: U.S. Government Printing Office, 1966.
3. Cottle, T. J. Going up; going down. *National Elementary Principal*, 1975, *54*, (4), 61, 59-62.
4. Ebel, R. L. The social consequences of educational testing. In A. Anatasi (Ed.), *Testing problems in perspective*. Washington, D.C.: American Council on Education, 1966.
5. Folds, G. H., Gazda, G. M. A comparison of effectiveness and efficiency of three methods of test interpretation. *Journal of Counseling Psychology*, 1966, *13*, 318-324.
6. Goslin, D. A. The social impact of testing. *Personnel and Guidance Journal*, 1967, *45*, 676-682.
7. Goslin, D. A. *The search for ability*. New York: Russell Sage Foundation, 1963.
8. Houts, P. L. Behind the call for test reform and abolition of the I.Q. *Phi Delta Kappan*, *57* (10), 669-673.
9. Kirkland, M. C. The effects of tests on students and schools. *Renewing Educational Research*, 1971, *41*, 303-350.
10. McCarthy, D. A study of the reliability of the Goodenough drawing test of intelligence. *Journal of Psychology*, 1944, *18*, 201-216.

11. Moss, H. A., Kagan, J. Stability of achievement in recognition selling behaviors from early childhood through adulthood. *Journal of Abnormal and Social Psychology*,1961, *62*, 504-513.

12. NAACP Report on minority testing. Published by the NAACP Special Contribution Fund, New York City, May, 1976.

13. Neulinger, J. Attitudes of American secondary school students toward the use of intelligence tests. *Personnel and Guidance Journal*, 1966, *45*, 337-341.

14. Sacks, E. L. Intelligence scores as a function of experimentally established social relationships between child and examiner. *Journal of Abnormal and Social Psychology*, 1952, *47*, 354-358.

15. Maryland State Department of Education. Improving student attitudes and skills for taking tests. An occasional paper on accountability. November 1975.

16. Tyler, B. B. Expectancy for eventual success as a factor in problem-solving behavior. *Journal of Educational Psychology*, 1958, *49*, 166-172.

Where Ignorance Is Bliss—'Tis Folly to Be Testing

ROBERT L. THORNDIKE
Professor Emeritus of Psychology and Education
Columbia University

There may have been a few of you who, when reading the title for my remarks, bristled slightly. "Who does this joker think he is," you may have said, "equating testing with being wise?" And I can't agree with you more! But the fault really lies in the old adage—because the antithesis of being ignorant is to be informed—not to be wise. Wisdom, like beauty, lies in the eye (or the cortex) of the beholder.

To be informed sounds, on the face of it, desirable—like baseball, apple pie and Chevrolet. But we need to ask: To what end does it profit us to be informed? And the uniform answer, it seems to me, is that we wish to be informed so that, being informed, we can make better decisions. Some folks may treasure information for its own sake, as others treasure bits of string, match books or rubber bands, but to most of us the fundamental value in being informed lies in the decisions that can be based on that information.

If that be so, our basic problem as makers of tests, peddlers of tests, or instructors in the use of tests is twofold. It is, first, to determine what information is useful in relation to what types of decisions and then to make it possible for the decider to get that information. It is, second, to try to bridge the gap from information to wisdom so that the relevant information is used with perceptiveness and restraint to lead to decisions that will foster growth, success and happiness for the individuals or groups concerned. My remarks today will be directed primarily to the first of these two problems, with the hope that there may be a little spinoff on the second.

It is important to recognize that there are a number of different types of decisions for which the information provided by testing may be relevant, and that the information needed for one type is likely to be quite different from that needed for another. The information needed by a teacher deciding whether to review capitalization of place names is

of quite a different sort from that needed by a twelfth grader trying to decide whether to apply for admission to Harvard. I would like to review some of these types with you and comment on the sort of information, and consequently the type of testing, especially achievement testing, that seems most appropriate for each.

First, let us turn our attention to instructional decisions. These are decisions, usually by the classroom teacher, of the type: "Mary knows what a prime number is. We don't need to teach her that, and can start her in on factoring." Or "Willie can't tell a complete sentence from a fragment. He needs help on this." A sound decision on whether to teach or not to teach topic B depends, in part at least, on information as to whether a student—or possibly, most of the students in a group—has mastery: first, of topic B itself, and second, of topics A_1, A_2, and so on that provide the foundations for topic B. If the student (or class) has already achieved a satisfactory level of mastery of B, to spend additional time teaching it seems a waste. On the other hand, if the student or class that cannot do B does not have command of certain of the A's, and these particular A's are *really* essential to learning B, to plunge into B without first mastering the A's seems likely to be an exercise in frustration and futility. This was the credo that motivated the authors who developed tests such as the Compass Diagnostic Arithmetic Tests back in the 1920s. And a revival of this credo appears to be what started the wave of enthusiasm for criterion-referenced tests in the past decade.

To the extent that aspects of the curriculum *are* sequential, to the extent that one *can* identify certain skills or certain bodies of knowledge that are necessary antecedents to successful study of other skills or bodies of knowledge and to the extent that one can define what constitutes an adequate level of mastery, this approach seems sound. But I believe that my "to the extent thats" represent very severe constraints upon the breadth of applicability of the "criterion-referenced" approach. It is no accident that most of the examples of criterion-referenced testing are drawn from arithmetic. Arithmetic is the academic subject that comes closest to comprising a sequential set of identifiable, discrete skills that can be fully mastered and in which later skills build upon the foundation of what has previously been learned. In primary reading, some of the basic word analysis and decoding skills may have similar status as essential contributors to fluent reading. And,

of course, there are numerous specific rules in language usage and elsewhere that represent teachable and testable specifics, even though they are not sequential in the sense that mastery of any one is essential to the teaching of any other. But much of school learning deals with material that is neither sequential nor organized in neat packages that can be fully mastered. What are the boundaries that define, and what constitutes mastery of reading comprehension or of the French Revolution? These represent broad domains—the one of skill, the other of knowledge—for which prerequisites or successors would be hard to specify and for which the concept of "mastery" at the 80 or 90 percent level seems to lose all meaning.

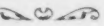
Even with fairly specific and definable skills, setting a standard of mastery can be a tricky business. Consider the rather precisely defined objective: When shown a 2-digit number, specifies whether or not it is a prime number. Relatively few in this room would assert that 25 or 88 are prime numbers, and the few who did would be persons with no conception at all or gross misconceptions of what a prime number is. However, my experience with previous groups like this indicates that a good many of you would unhesitatingly identify 51 or 91 as being prime —though of course they aren't. Here, as in many other cases, it makes a world of difference which exemplars one chooses in order to test mastery of even a sharply delimited skill domain, and for many students, whether one will or will not conclude that they have achieved mastery in terms of some specified proportion of successes will depend critically upon the specific tasks that have been chosen to exemplify the domain.

As a minor detour, it seems to me that from a psychometric point of view assessment of real *mastery* is most efficiently achieved by using tasks that represent the more difficult exemplars of the domain, so long as they do not introduce other and irrelevant sources of difficulty. We test prime number mastery better with 51 than with 50, mastery of the basic addition combinations better with 7+8 than with 2+3. Success on the easy items tells very little about mastery, though it may signify a good beginning; success on the hardest items tells a lot.

Another point to be borne in mind is that the performance of a learner who is just picking up a new competence tends to fluctuate from day to day and week to week. One doctoral student with whom I worked studying foreign students who were learning English, and using a set of mini-tests of specific English usages, found markedly lower consistency between two tests given only a week apart than within the items of a single test, the respective reliability coefficients for 4-item

tests being respectively about .80 and .60. Two separated short tests to appraise mastery should permit a wiser decision than a single test twice as long.

Thus, criterion-referenced tests built of the critical examples of a defined skill, perhaps repeated to check upon stability of mastery, can in certain limited areas provide the information basic to wise instructional decisions.



But a wide range of other decisions that arise in the process of education call for information on the performance of individuals and of groups. We can distinguish selection decisions, placement decisions, decisions involving curricular choice and resource allocation, and a whole set of decisions that we might call guidance decisions or personal decisions. What sorts of information provided by what sorts of testing instruments will permit decisions of these kinds to be made more wisely? Let us turn our attention for a bit to selection decisions.

Implicit in the very concept of selection is a situation in which there are more aspirants to some particular good, be it admission to a program in veterinary medicine, a berth on the Dallas Cowboys, or an executive secretary's job with the president of Widgets International, than there are positions to be filled. There is the often painful task of choosing among persons all of whom may be at least minimally qualified, trying to pick the best, or at least the better qualified from among the applicants. The regression of some index of job performance upon score on a predictor test represents one type of information to guide such a decision.

We have in the past tended to view the selection enterprise in terms analogous to the economist's cost-benefit analysis. More efficient employees can be considered to generate benefits to the employer in improved productivity, at whatever cost is involved in a recruitment and testing program. But in education we dare not take quite as narrow a view of costs and benefits as might be acceptable for the professional football coach, or the industrial personnel manager. The benefits cannot simply be represented by grade point average, but need to take account of the broader utility of the person in the larger society. An adequate medical student who will provide service in the urban ghetto or the rural South may represent greater social utility than a brilliant one who will compete for patronage in a middle-class suburb.

Decisions involve not only those facts that testing can supply, but also a value system that has nothing to do with tests and testing. This is the fundamental consideration that has been back of much of the debate of the past decade about "fair testing" even though it has seldom been identified as such. Competing definitions of what is "fair" differ primarily not on psychometric issues but on the question of whose utility is paramount. The classical approach that used tests and any other available information to establish for each individual a predicted academic or job performance, and then selected those for whom the prediction was highest, adopted a view narrowly focused on the employer's or selector's utility. This narrow view may be acceptable in the football coach whose values must focus solely on winning as many games as possible. It becomes more questionable in an employer whose decisions structure the job opportunities for large segments of our society, and still more questionable in the admissions office of an educational institution that exists only to serve society. Utility in a graduate from a college or professional school must be viewed not solely or primarily in terms of grade point average nor of income, X years out of college, but primarily in the broader sense of value to society. This is, of course, a fuzzy, ambiguous notion, and there will be wide differences in perception of where the common good lies. But unless we can achieve consensus on such value questions, no amount of psychometric elegance or refinement will bring us to agreement. It is important, I believe, that we recognize that this is where the shoe pinches. Perhaps we can develop a calculus of values that will permit us to specify our utilities, and to clarify our differences in the utility that we attach to different outcomes, but for the present such a calculus seems quite a remote prospect. And even clarification will not guarantee agreement.

However *one* element in any judgment about utility *is* the probability that the candidate will perform satisfactorily in the tasks to which he seeks acceptance. How well will this candidate master the mysteries of torts or the skills of operating a Selectric typewriter? And what model of test will provide information that will be useful in indicating the performance that we can expect from candidates X, Y and Z? I submit that it is likely to be some general assessment of a broad area of knowledge or skill. With what speed and understanding does Candidate X read social studies material—not, what is his mastery of the economic geography of Brazil. How well does the secretarial candidate spell a broadly representative sample of words—not has he or she

mastery of the ei/ie rules (and their several exceptions). A good old-line survey test, with expectancy tables that indicate probable criterion performance at each score level will permit us to be more usefully informed on the individual's prospects for effective performance than will a narrowly focused criterion-referenced mastery test of some highly specific skill.

The same thing is true, I believe, for a wide range of placement, guidance and personal decisions. The type of information that could be useful in deciding whether a freshman would be likely to learn more in the remedial English section, the regular course, or a special course in literature or in writing would be a broad appraisal of writing skills, of competence in reading literary material, or conceivably of knowledge of grammar and syntax rather than a focused mastery test of use of the semicolon or of agreement between subject and verb. A personal decision to apply to Harvard would be more soundly based on a broad survey measure of high school achievement with performance compared to norms for other high school juniors than on a mastery chemistry test on the periodic table.

Even decisions relating to curricular modifications or resource allocation would seem to call primarily for broad appraisals of the comprehensive set of objectives that the school system is trying to achieve. As a matter of fact, there is little case being made for the narrowly focused criterion-referenced mastery test as a basis for curricular or resource allocation decisions. The current watchword here seems to be "objective-referenced." This appears to mean that the school system states in detail, with a good deal of specificity and usually at great length, just what its instructional goals are, and that each test exercise is designed to assess some one of those objectives. One can hardly quarrel with a test design in which the test exercises are built to match the content and process objectives that seem important as goals of schooling. Every achievement test worth its salt has always been built around a blueprint of curricular objectives. The question would seem to be whether the objectives of schooling are sufficiently different from one school system to another for it to be desirable to prepare a separate and unique array of test exercises for each.

Undoubtedly there are instances in which objectives are local and idiosyncratic. When a social studies program focuses on local history or local economic geography, for example, New York, Illinois, and California will need completely distinct evaluation instruments. When particular state or local curricula operate with quite distinctive se-

quences for the presentation of topics, a common appraisal may make sense only at the end when all have arrived at nearly the same final destination. In some areas, at least, it will be unreasonable to expect children to have learned what they haven't been taught.

However, development and use of special testing instruments for local situations is not without its costs, and these costs lie not solely in the hours and dollars required to develop and print the special tests. Though it will be possible to determine what proportion of children in a given school system succeed on a specific test item or group of test items, this proportion will vary from low to high depending upon the basic difficulty of the item, in addition, to some extent, to its relationship to what has been taught and emphasized, and it will be difficult to know whether one should be pleased or distressed by the percentages. Unless test exercises are limited to the simplest exemplars of the minimum essentials—in which case they will give a very incomplete picture of the full range of learning sought and to a degree achieved by the school—there will be varying proportions of children who will not be able to do an item. If the item tests the limits of skill or knowledge, the proportion who cannot manage it may be quite high. Excepting as the items have been drawn from nationally standardized tests for which item norms have been developed based on a representative sample of school children, there will be no meaningful external basis for comparison. It will be difficult, if not impossible, to determine whether high and low percentages of right answers are to be attributed to the successes and failures of the program or to the inherent ease or difficulty of the test exercises. Furthermore, it will be impossible to determine at what cost, in achievement of content and skills omitted from the local assessment instruments, any gains in the objectives assessed in those instruments have been achieved. It will, of course, be possible to make internal comparisons—between communities within a state, between schools within a community, between classes and pupils within a school. And these may be the comparisons that are relevant for decisions concerning resource allocation, concerning local shifts of emphasis or local remedial effort. There is clearly likely to be some advantage in having these internal comparisons based on locally shared and agreed-upon objectives. The issue is whether the gain is worth the cost.

Some curricular and resource allocation decisions clearly call for fine-grained information at the level of the item or the short subset of items. Whether additional instruction on prime numbers is needed in the

5th grade can best be judged by knowing what percent of 5th graders think that 15 or 27 or 91 are prime numbers. And whether one school needs to provide special help on identifying the main idea of a paragraph depends on whether students in that school show noticeably lower proportions of success on exercises requiring that skill than do students in other schools that recruit from a similar student population. Norms at the item level, which have been rather generally provided by test publishers during the past decade, provide valuable information in terms of which to make such judgments and decisions. We can expect that in the future publishers will continue to provide normative information not only on test scores but upon items. But for broader assessments of relative success on the major segments within a skill or between skills, normative information on test scores will continue to be needed.

Turning away from achievement tests, I would assert that microanalyses of successes and failures on specific test items make essentially no sense on tests developed to measure aspects of aptitude, as contrasted with measures of achievements related to specific aspects of an educational program. I am talking here about analyses used as a basis for decisions about persons, and not about decisions on the development and construction of tests. Obviously, item analysis plays a central role in aptitude test construction. To know that on the Wechsler Intelligence Scale for Children (WISC) a 14-year-old got a full-scale IQ of 110 tells us something potentially meaningful about that child's probable success in an algebra class. To know that the child got 14 of the 18 items on the arithmetic test right may also be a useful datum. But to know the one specific fact that he got the correct answer on "36 dollars at 4 dollars an hour" is of minimal help in our appraisal either of his general scholastic aptitude, of his more specific quantitative ability, or of his likelihood of being a successful algebra student. Aptitudes represent general areas of competence that have no precise lateral boundaries and no upper limits. We appraise them by sampling broadly from some extended and ill-defined domain, often relating performance to that of others, since our inferences are predictive, and most of our predictions are inherently relative rather than absolute. I find it almost impossible to conceive how microanalysis of single aptitude test items would contribute anything useful to decisions by or about persons.

Wisdom in relation to test scores calls for information on the predictive significance of the test scores in many specific contexts. There is a big gap between knowledge in general of the validity of Scholastic Aptitude Test (SAT) scores for predicting college success, even when that success is narrowly defined as freshman grade point average, and knowledge of the proportion of the applicants with SAT-V scores of 450 who are admitted to Siwash, and the distribution of GPAs of those persons when they get there. The College Entrance Examination Board, the American College Testing Program, and various of the state testing service groups have steadily increased their efforts to make this type of institution-specific information accessible beyond the smoke-filled offices of admissions directors not only to school guidance staffs but to the individual students who, in the final analysis, must make decisions about their own futures.

A certain reluctance on the part of some institutions to make the information available is understandable. There is an element of self-fulfilling prophecy in letting information about one's institutional past structure one's institutional future. But this *is* the type of information that is most directly relevant to decisions about whether or where to apply for admission. In all the settings in which test results are used for guidance decisions or personal decisions, improved communication systems are needed for assembling and transmitting specific information on the implications of those test results for the alternative educational or vocational choices that are being faced.

This concern points also to a basic problem that we always face when we try to base selection or counseling or personal decisions upon the data that we have meticulously collected. Inevitably, these are data from the past—sometimes from the fairly remote past, yet we use them for decisions that relate to the future—sometimes the fairly remote future. For example, Project Talent's *Career Data Book*[1] reports in 1973 the sorts of students tested in 1960 who were in various occupational categories five years after the year in which they were or would have been in the twelfth grade. The counselor in 1976 who uses these data is helping students to make decisions that will be operational in the 1980s. These can be wise decisions only to the extent that occupational opportunities and demands of 1980 match those of 1965 to 1969 when Talent's students were making their occupational choices. The assumption may be reasonable; the world changes fairly slowly. It is certainly nec-

essary. The only way we can anticipate the future is by knowing the past. But it should also be recognized.

Just as it is necessary to use data from the past to make inferences and decisions about the future, and to assume a continuity of the standards, conditions and relationships of the past into the future, so it is also necessary to project relationships from one specific setting to another. It is manifestly impossible to replicate empirical validation studies in every plant, office or school in which a testing procedure might be used. Time, numbers, availability of sound performance estimates, as well as financial resources, all set limits on what can be done. So we must often use findings from other plants, offices or schools and apply them to our present context.

Yet our skills of specifying the dimensions of similarity and difference between jobs in different settings or intellectual demands in different programs constitute a serious limitation on the confidence with which we can generalize relationships of predictors to performance, and standards of acceptable performance in different settings. There have been calls, within the field of vocational psychology, for studies of the microstructure of jobs, and of the relationships of test scores to elements of that microstructure. I am not aware that we have made great strides in that direction, and I am not sure what the payoff will be.

But if we are to generalize with any confidence from one academic or job setting to another, it may well be that some more specific analyses of just what it is in a job that is predicted by our test scores—or other items of information about a person, so far as that is concerned—will be essential. In the interim, we can only maintain a discreet tentativeness in our generalization of data to new situations, trying as best we can to assess the degree of identity between the setting of available data and the setting to which we would apply them.

It is, alas, no easy matter to translate information to wisdom. Facts are not simple but complex, and values are not uniform but diverse. We may, to quote another aphorism, conclude with Pope—Alexander, not Paul—that "a little learning is a dangerous thing." We may conclude, as some groups and organizations appear to have done, that it is better to forego information about the achievements and abilities of our students—individually and collectively—because of the possibility that we may use that information unwisely. We may abandon the attempt to understand better and to teach others better to understand the implications of test scores. We may elect to remain blissfully ignorant of the information that tests can give in the hopes that thus we can

avoid folly. Or we may continue the struggle to understand, to appreciate, and to be wise.

Reference

1. Flanagan, J. C. *Career data book: Results from Project Talent's 5-year follow-up study.* Palo Alto, CA: American Institutes for Research, 1973.